REMARKS

In the subject Office Action, the Examiner rejected claims 22-27 under 35 U.S.C. § 101 and the first paragraph of Section 112 for lack of utility and enablement. The Examiner also rejected claims 22 and 27 under the second paragraph of 35 U.S.C. § 112 as being indefinite. The Examiner maintained and held in abeyance the priority determination. Lastly, the Examiner maintained the rejection of claims 22-27 under 35 U.S.C. § 102 as being anticipated by Holtzman et al.

Applicants respectfully traverse the rejections and request that the Examiner consider the following remarks in response to the Office Action.

Claim 27 has been cancelled. Claim 22 has been amended to clarify that the claimed antibody specifically binds the polypeptide of SEQ ID NO:69. Support for amendment of claim 22 may be found throughout the specification, including on page 16, lines 1-3. Claims 22-26 are now pending.

Priority Determination:

Applicants note that in the Office Action dated March 12, 2003, the Examiner agreed to delay the determination of priority until after the utility rejection is fully resolved. Applicants believe that they have resolved this utility issue and therefore direct the Examiner's attention to the arguments made herein, previous Applicant's arguments submitted on January 30, 2003, and the Goddard Declaration, also filed January 30, 2003. Applicants submit that this evidence demonstrates that Applicants are entitled to priority of at least December 17, 1997 **(correct?).** Therefore, Applicants request reconsideration of the determination of priority in view of the submission of all the evidence showing utility of the claimed invention.

Rejection of Claims Under 35 U.S.C. § 101 and 112, First Paragraph, Lack of Utility and Enablement:

The Examiner has rejected claims 22-27 under both 35 U.S.C. § 101 and 112, first paragraph, as being drawn to an invention that lacks utility. More specifically, the Examiner stated that the invention is not supported by either a credible, specific and substantial utility or a well established utility. Furthermore, the Examiner indicated that the Goddard Declaration, filed under CFR 132 on January 30, 2003, is insufficient to overcome the rejection because, even though it demonstrates that the increase of the DNA levels of certain markers may be indicative of cancer, the Declaration does not address how the DNA levels relate to the protein expression levels. In fact, the Examiner argues that an increase in the mRNA level expression does not necessarily result in an increased protein expression levels. Applicants respectfully request reconsideration of the rejection of claims 22-27, 30, 31, 33, and 34 for the reasons discussed below.

Applicants respectfully direct the Examiner's attention to several publications and abstracts of publications that demonstrate that mRNA levels correlate with protein expression levels, attached hereto as Appendix A. These publications make it clear that skilled artisans recognize that the expression levels of mRNA often correlate with the protein expression levels. For example, Maruyama *et al.* (Am. J. of Pathol., Sept. 1999, Vol. 155, No. 3, pgs. 815-822) showed a correlation between mRNA levels and protein levels of three of the helix-loop-helix proteins, Id-1, Id-2, and Id-3. According to Maruyama *et al.*, the mRNA and protein levels of all three species were increased in pancreatic cancer tissues as compared to the normal or chronic pancreatitis control tissues. Also, Ginestier *et al.* (Am. J. Pathol., Oct. 2002, 161(4):1223-33) demonstrated a correlation between cDNA (cDNA array analysis) and protein expression levels (using tumor tissue microarray analysis) in one-third of the examined molecules with proven or suspected role in breast cancer.

Applicants also include several other publications for the Examiner's consideration. For example, Dalifard *et al.* (Int. J. Mol. Med., May 1998, 1(5):855-61) showed that in breast cancer, a correlation (r=0.85) existed between c-erbB2 (oncogene encoding for p185 protein) expression (as determined by Southern blot method) and p185 protein expression levels (as determined by immunoenzymetric assay). Also, Hareuveni *et al.* (Eur. J. Biochem., May 1990, 189(3): 475-86) found a correlation between expressed tumor antigen species with the allelic forms as well as significantly increased protein expression in breast cancer. Furthermore, Barr *et al.* (J. Parasitol., April 2003, 89(2):381-4) demonstrated that in a canine model of Chagas disease, mRNA levels (as determined by Northern blotting) and protein expression levels (as determined by Western blotting) of the plasma membrane calcium pump (PMCA) were increased in cardiac tissue by 77% and 58%, respectively, as compared to normal controls.

Accordingly, because the RNA levels and the protein expression levels have been found to correlate in different types of cancers as well as other diseases, Applicants assert that it is reasonable to expect the protein levels of the polypeptide encoded by the SEQ ID NO: 69 to be increased in cancer tissue. The increase in protein levels can then be detected by the antibodies of this invention.

Considering these remarks, Applicants respectfully assert that the claimed invention has utility and is fully enabled. Accordingly, Applicants request that the Examiner reconsider and withdraw the rejections under § 101 and the first paragraph of § 112.

Rejection of Claims Under 35 USC § 112, Second Paragraph - Indefiniteness:

The Examiner rejected claims 22 and 27 under 35 U.S.C. § 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which Applicants regard as the invention. Specifically, the Examiner acknowledges that the difference between claims 22 and 27 is the use of the terms "binds" and "specifically binds," respectively. However, the Examiner asserts that the specification does not define these terms and that Applicants do not support their contention that a skilled artisan would recognize the definitions submitted by the Applicants. Applicants respectfully disagree. Claim 22 now recites "an antibody that specifically binds to the polypeptide shown in Figure 26 (SEQ ID NO:69)," and claim 27 has been cancelled. Therefore, the rejection is moot. Applicants maintain that one skilled in the art would recognize the definition of "specifically binds," which is also found throughout the specification, including on page 16, lines 1-3. Accordingly, Applicants request that the Examiner's rejection of claims 22 and 27 under § 112, second paragraph be withdrawn.

Rejection of Claims Under 35 USC § 102 Anticipation:

Claims 22-27 remain rejected under 35 U.S.C. § 102(a) and (e) as being anticipated by U.S. Patent Number 6,225,085 (Holtzman *et al.*).

The Examiner notes that Holtzman *et al.* disclose a polypeptide, LRSG, which is 98.4% identical to the SEQ ID NO: 69 of the instant application. The Examiner notes that although Holtzman *et al.* disclose LRSG, which differs from SEQ ID NO: 69 of the present application by 75 residues, "at 598 out of 673 [amino acids], the two proteins are absolutely identical."

Applicants submit that SEQ ID NO: 69 of Applicants invention is not identical to the sequence disclosed in Holtzman *et al.* and therefore claims 22-27 are not anticipated. According to the MPEP § 2131, "a claim is anticipated only if each and every element as set forth in the claim is found, either expressly or inherently described, in a single prior art reference." *Verdegaal Bros. v. Union Oil Co. of CA*, 814 F.2d 628, 631, 2 USPQ2d 1051, 1053 (Fed. Cir. 1987). "The *identical* invention must be shown in as complete detail as is contained in the ... claim." *Id.* Also, *see* MPEP § 2131; *Verdegaal Bros. v. Union Oil Co. of CA*, 814 F.2d 628, 631, 2 USPQ2d 1051, 1053 (Fed. Cir. 1987). Applicants submit that two amino acid sequences that differ by 75 amino acids are not identical.

The Examiner indicated that the sequence differences existing between the polypeptide of this invention and the polypeptide sequence according to Holtzman *et al.*

do not correspond to major structural features that would affect the folding as well the ability of antibodies to bind to these polypeptides, specifically or otherwise. However, Applicants direct the Examiner attention to a review by Bowie *et al.* (Science, March 1990, 247:1306-1310) illustrating several examples of how a single amino acid substitution in a polypeptide sequence has significant structural and functional consequences. To carry out their function, proteins require the binding residues or the catalytic residues to be precisely oriented in three dimensions (Bowie *et al.*, Science, March 1990, 247:1306-1310). Therefore, due to the differences of 75 residues in two amino acid sequences, these two amino acid sequences likely encode two different, *i.e.* not identical polypeptides.

More specifically, as previously noted, the polypeptide described by Holtzman *et al.* contains an additional 75 residues not present in the DNA44804 polypeptide of the present invention. Applicants respectfully point out that this deletion of 75 residues from the polypeptide of this invention will have significant effects on the structure, stability and unfolding-refolding of the polypeptide. For example, Chaudhuri *et al.* (J. Mol. Biol. 1999, 285(3):1179-94) have shown that although the overall structure of the two proteins, the native and recombinant proteins, might be the same, the extra methionine residue at the N terminus of the recombinant protein remarkably affected the native-state stability and electric properties. According to Chaudhuri *et al.*, the protein with one less residue was more stable, was less negatively charged and unfolded faster then the protein with the additional residue. This author concluded that the difference in one residue resulted in electrostatic interactions and destabilization of the protein containing the additional residue through a conformational entropy effect. Based on these findings, Applicants submit that the deletion of 75 amino acids would have significant consequences on the structure, stability and the ability of the polypeptide to bind to the antibodies of this invention. As a result, it is clear that the polypeptide taught by Holtzman *et al.* does not anticipate the claimed invention. Accordingly, the Holtzman *et al.* reference is not a proper anticipatory reference of any currently pending claims. The Applicants respectfully request that the Examiner reconsider and withdraw this rejection of the claims.

The Applicants respectfully assert that the application is now in condition for allowance and request a timely notice of allowance be issued in this case. Should the Examiner feel a discussion would expedite the prosecution of this application, the Examiner is kindly invited to contact the undersigned.

Applicants believe no fee is due in connection with the filing of this Amendment, however, should any fees be deemed necessary for any reason relating to this paper, the Commissioner is hereby authorized to deduct said fees from Brinks Hofer Gilson & Lione Deposit Account No. 23-1925. a duplicate copy of this document is enclosed.

Respectfully submitted,

C. Noel Kaman

C. Noel Kaman
Registration No. 51,857
Attorney for Applicant

BRINKS HOFER GILSON & LIONE
P.O. BOX 10395
CHICAGO, ILLINOIS 60610
(312) 321-4200

Appendix A.

American Journal of Pathology, Vol. 155, No. 3, September 1999
Copyright © American Society for Investigative Pathology

Id-1 and Id-2 Are Overexpressed in Pancreatic Cancer and in Dysplastic Lesions in Chronic Pancreatitis

Haruhisa Maruyama,* Jörg Kleeff,* Stefan Wildi,* Helmut Friess,† Markus W. Büchler,† Mark A. Israel,‡ and Murray Korc*

From the Division of Endocrinology, Diabetes, and Metabolism, Departments of Medicine, Biological Chemistry and Pharmacology, University of California, Irvine, California; the Department of Visceral and Transplantation Surgery,† University of Bern, Bern, Switzerland; and the Preuss Laboratory,‡ Department of Neurological Surgery, University of California, San Francisco, California*

Id proteins antagonize basic helix-loop-helix proteins, inhibit differentiation, and enhance cell proliferation. In this study we compared the expression of Id-1, Id-2, and Id-3 in the normal pancreas, in pancreatic cancer, and in chronic pancreatitis (CP). Northern blot analysis demonstrated that all three Id mRNA species were expressed at high levels in pancreatic cancer samples by comparison with normal or CP samples. Pancreatic cancer cell lines frequently coexpressed all three Ids, exhibiting a good correlation between Id mRNA and protein levels, as determined by immunoblotting with highly specific anti-Id antibodies. Immunohistochemistry using these antibodies demonstrated the presence of faint Id-1 and Id-2 immunostaining in pancreatic ductal cells in the normal pancreas, whereas Id-3 immunoreactivity ranged from weak to strong. In the cancer tissues, many of the cancer cells exhibited abundant Id-1, Id-2, and Id-3 immunoreactivity. Scoring on the basis of percentage of positive cells and intensity of immunostaining indicated that Id-1 and Id-2 were increased significantly in the cancer cells by comparison with the respective controls. Mild to moderate Id immunoreactivity was also seen in the ductal cells in the CP-like areas adjacent to these cells and in the ductal cells of small and interlobular ducts in CP. In contrast, in dysplastic and atypical papillary ducts in CP, Id-1 and Id-2 immunoreactivity was as significantly elevated as in the cancer cells. These findings suggest that increased Id expression may be associated with enhanced proliferative potential of pancreatic cancer cells and of proliferating or dysplastic ductal cells in CP. *(Am J Pathol 1999, 155:815–822)*

Basic helix-loop-helix (bHLH) proteins play an important role as transcription factors in cellular development, proliferation, and differentiation.[1,2] The basic domain of the bHLHs is required for binding to an E-box DNA sequence, thus promoting transcription of specific target genes. The HLH domain promotes dimer formation with various members of the bHLH protein family.[1,2] Homodimers of the class B family of bHLH proteins, including *Myo*D, *Neuro*D, and numerous other proteins, are known to activate tissue-specific genes.[3–5] These tissue-specific bHLHs typically form heterodimers with widely expressed class A bHLHs, which include proteins encoded by E2A, E2–2, HEB, and other genes (also termed E-proteins).[6–9] These heterodimers activate transcription of genes that are associated with differentiation.

Id genes encode a family of four HLH proteins that lack the basic DNA binding domain.[1,10] They act as dominant-negative HLH proteins by forming high affinity heterodimers with other bHLH proteins, thereby preventing them from binding to DNA and inhibiting transcription of differentiation-associated genes.[10–12] Id gene expression is down-regulated on differentiation in many cell types *in vitro* and *in vivo*.[13–18] In addition, Id proteins seem to be required for cell cycle progression through G_1/S phase in certain cell types, and interaction between Id-2 and pRB is associated with enhanced proliferation in some cell lines *in vitro*.[19–23]

Pancreatic cancer is the fifth leading cause of cancer death in the United States, with a mortality rate that virtually equals its incidence rate.[24] This malignancy is often associated with the overexpression of a variety of mitogenic growth factors and their receptors, and by oncogenic mutations of K-*ras* and inactivation of the p53 tumor suppressor gene.[25] We have recently reported that pancreatic cancers overexpress the HLH protein Id-2, and that enhanced expression of this protein is evident in the cytoplasm of the cancer cells within the pancreatic tumor mass.[26] It is not known, however, whether the expression of other Id proteins is altered in this malignancy, or whether their expression is altered in chronic pancreatitis

Contract grant sponsor: National Cancer Institute. Contract grant number: U. S. Public Health Service grant CA-40162.

Accepted for publication May 24, 1999.

Address reprint requests to Dr. Murray Korc, Division of Endocrinology, Diabetes and Metabolism, Medical Sciences I, C240, University of California, Irvine, CA 92697. E-mail: mkorc@uci.edu.

(CP), an inflammatory disease that is characterized by dysplastic ducts, foci of proliferating ductal cells, acinar cell degeneration, and fibrosis.[27] We now report that there is a five- to sixfold increase in Id-1 and Id-2 mRNA levels and a twofold increase in Id-3 mRNA levels in pancreatic cancer by comparison with the normal pancreas. In contrast, overall Id mRNA levels are not increased in CP.

Patients and Methods

Normal human pancreatic tissue samples from 7 male and 5 female donors (median age 41.8 years, range 14–68 years), CP tissues from 13 males and 1 female (median age 42.1 years; range 30–56 years), and pancreatic cancer tissues from 10 male and 6 female donors (median age 62.6 years; range 53–83 years) were obtained through an organ donor program and from surgical specimens from patients with severe symptomatic chronic pancreatitis or pancreatic cancer. A partial duodenopancreatectomy (Whipple/pylorus-preserving Whipple; $n = 13$), a left resection of the pancreas ($n = 2$), or a total pancreatectomy ($n = 1$) were carried out in the pancreatic cancer patients. According to the TNM classification of the Union Internationale Contre le Cancer (UICC) 6 tumors were stage 1, 1 was stage 2, and 9 were stage 3 ductal cell adenocarcinoma. Freshly removed tissue samples were fixed in 10% formaldehyde solution for 12 to 24 hours and paraffin-embedded for histological analysis. In addition, tissue samples were frozen in liquid nitrogen immediately on surgical removal and maintained in −80°C until use for RNA extraction. All studies were approved by the Ethics Committee of the University of Bern, Bern, Switzerland, and by the Human Subjects Committee at the University of California, Irvine, California.

Northern Blot Analysis

Northern blot analysis was carried out as described previously.[26,28] Briefly, total RNA was extracted by the single step acid guanidinium thiocyanate phenol chloroform method. RNA was size-fractionated on 1.2% agarose/1.8 mol/L formaldehyde gels, electrotransferred onto nylon membranes, and cross-linked by UV irradiation. Blots were prehybridized and hybridized with cDNA probes and washed under high stringency conditions. The following cDNA probes were used: a 979-bp human Id-1 cDNA probe, a 440-bp human Id-2 cDNA probe, and a 450-bp human Id-3 cDNA probe, covering the entire coding regions of Id-1, Id-2, and Id-3, respectively. A *Bam*HI 190-bp fragment of mouse 7S cDNA that hybridizes with human cytoplasmic RNA was used to confirm equal RNA loading and transfer. Blots were then exposed at −80°C to Kodak BioMax-MS films and the resulting autoradiographs were scanned to quantify the intensity of the radiographic bands.[26,28] For each sample the ratio of Id mRNA expression to 7S expression was calculated. To compare the relative increase in expression of the respective Id mRNA species in the cancer and CP samples, the same normal samples were used for normal/cancer and normal/CP membranes. The median score for Id-1, Id-2, and Id-3 mRNA levels in these normal samples was set to 100. Statistical analysis was performed with SigmaStat software (Jandel Scientific, San Raphael, CA). The rank sum test was used, and $P < 0.05$ was taken as the level of significance.

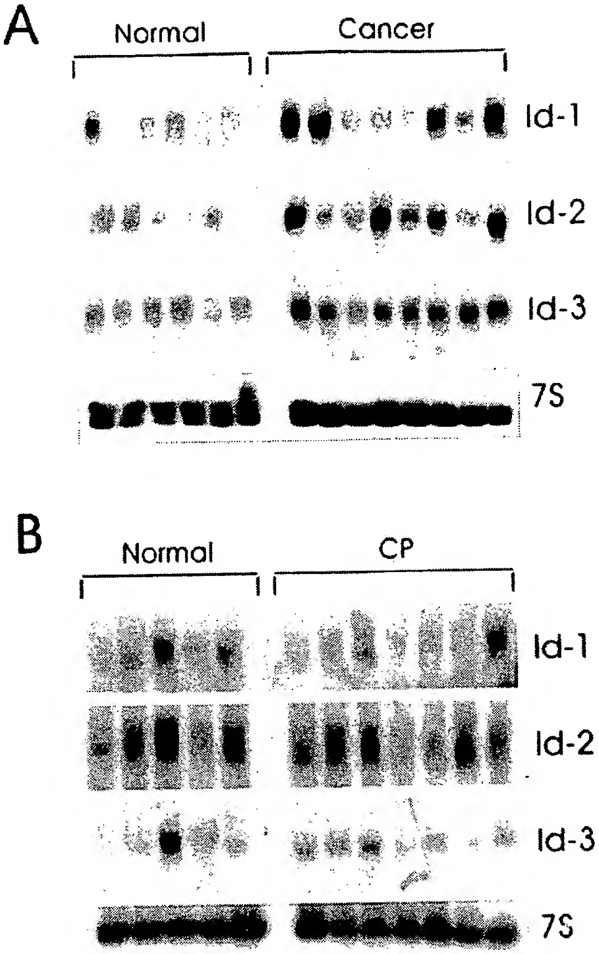


Figure 1. mRNA expression of Id-1, Id-2, and Id-3 in pancreatic cancer and chronic pancreatitis. Total RNA (20 μg/lane) from six normal, eight cancerous, and seven chronic pancreatitis tissue samples were subjected to Northern blot analysis using ^{32}P-labeled cDNA probes (500,000 cpm/ml) specific for Id-1, Id-2, and Id-3, respectively. A 7S cDNA probe (50,000 cpm/ml) was used as a loading and transfer control. Exposure times of the normal/cancer blots were 1 day for all Id probes, and 2 days for the normal/CP blots. Exposure time was 4 hours for mouse 7S cDNA. By comparison with the normal samples, Id-1 and Id-3 mRNA levels were elevated in 8 and 9 cancer samples, respectively, whereas Id-2 was elevated in 6 cancer samples.

Cell Culture and Western Blot Analysis

PANC-1, MIA-PaCa-2, ASPC-1, and CAPAN-1 human pancreatic cell lines were obtained from ATCC (Manassas, VA). COLO-357 human pancreatic cells were a gift from Dr. R. S. Metzger (Durham, NC). Cells were routinely grown in DMEM (COLO-357, MIA-PaCa-2, PANC-1) or RPMI (ASPC-1, CAPAN-1) supplemented with 10% fetal bovine serum, 100 U/ml penicillin, and 100 μg/ml streptomycin. For immunoblot analysis, exponentially growing

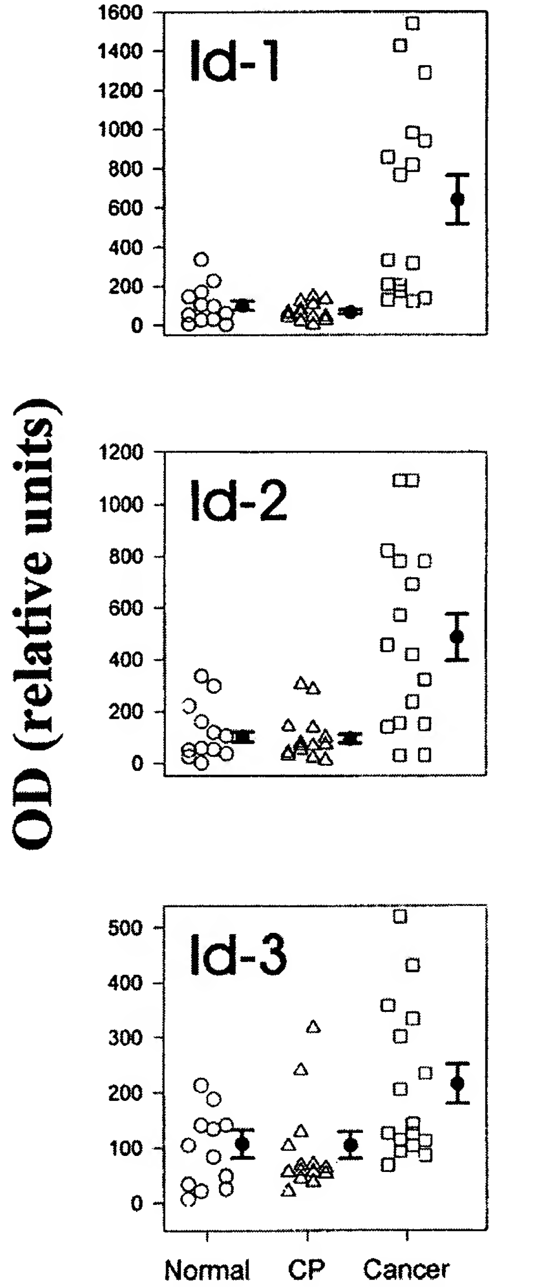


Figure 2. Densitometric analysis of Northern blots. Autoradiographs of Northern blots from 12 normal, 14 CP, and 16 pancreatic cancers were analyzed by densitometry. mRNA levels were determined by calculating the ratio of the optical density for the respective Id mRNA species in relation to the optical density of mouse 7S cDNA. To compare the relative increase in expression of the respective Id mRNA species in the cancer and CP samples, the same normal samples were used for normal/cancer and normal/CP membranes. Normal pancreatic tissues are indicated by circles, CP tissues by triangles, and cancer tissues by squares. Data are expressed as median scores ± SD. By comparison with the normal samples, only the cancer samples exhibited significant increases: 6.5-fold ($P < 0.01$) for Id-1, fivefold ($P < 0.01$) for Id-2, and twofold ($P = 0.027$) for Id-3.

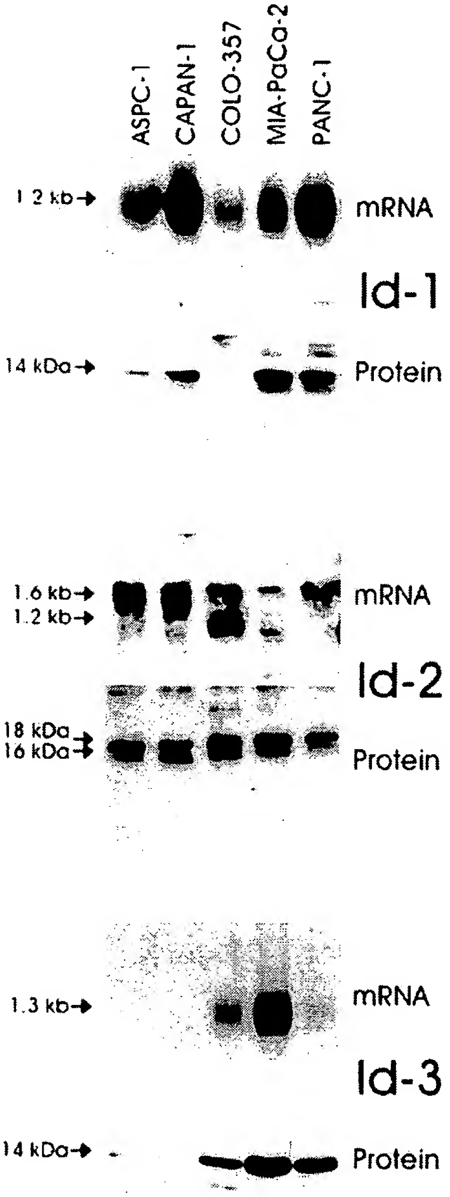


Figure 3. Id mRNA and protein expression in pancreatic cancer cell lines. Upper panels: Total RNA (20 μg/lane) from 5 pancreatic cancer cell lines were subjected to Northern blot analysis using ^{32}P-labeled cDNA probes (500,000 cpm/ml) specific for Id-1, Id-2, and Id-3, respectively. Exposure times were 1 day for all Id probes. Lower panels: Immunoblotting. Cell lysates (30 μg/lane) were subjected to SDS-PAGE. Membranes were probed with specific Id-1, Id-2, and Id-3 antibodies. Visualization was performed by enhanced chemiluminescence.

cells (60–70% confluent) were solubilized in lysis buffer containing 50 mmol/L Tris-HCl, pH 7.4, 150 mmol/L NaCl, 1 mmol/L EDTA, 1 μg/ml pepstatin A, 1 mmol/L phenylmethylsulfonyl fluoride (PMSF), and 1% Triton X-100. Proteins were subjected to sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), transferred to Immobilon P membranes, and incubated for 90 minutes with the indicated antibodies and for 60 minutes with secondary antibodies against rabbit IgG. Visualization was performed by enhanced chemiluminescence.

Immunohistochemistry

Specific rabbit anti-human Id-1 (C-20), Id-2 (C-20), and Id-3 (C-20; all from Santa Cruz Biotechnology, Santa

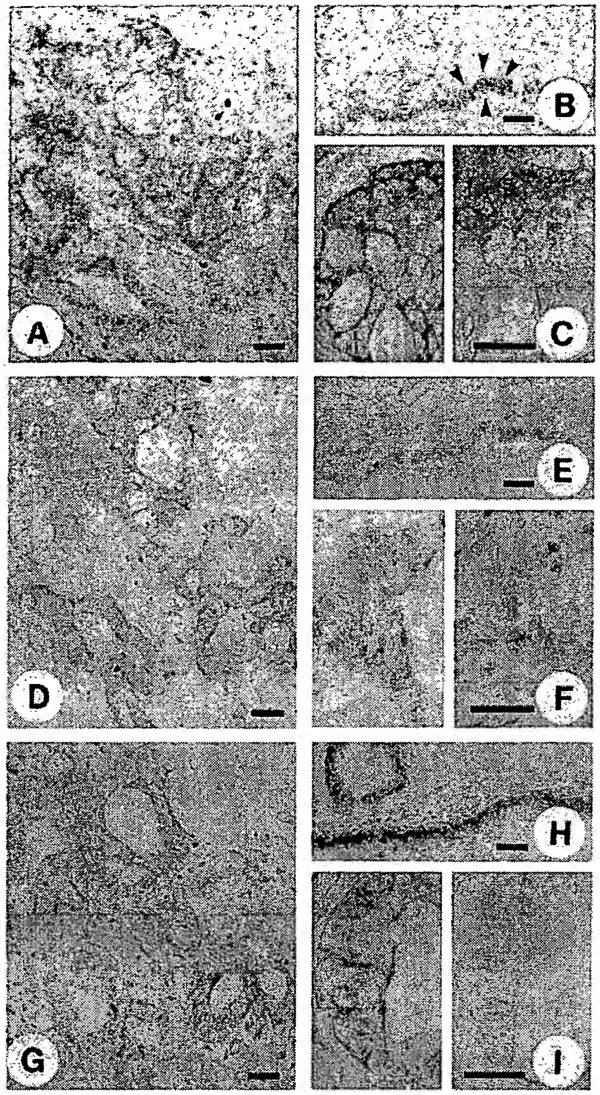


Figure 4. Normal and cancerous pancreatic tissues were subjected to immunostaining using highly specific anti-Id-1 (A-C), anti-Id-2 (D-F), and anti-Id-3 (G-I) antibodies as described in the Methods section. Moderate to strong Id-1 immunoreactivity was present in the cytoplasm of duct-like cancer cells (A and C, left panel). In the normal pancreas there was weak Id-1 immunoreactivity in the ductal cells (B). Preabsorption with the Id-1-specific blocking peptide abolished the Id-1 immunoreactivity (C, right panel). Strong Id-2 immunoreactivity was observed in the cytoplasm of the cancer cells that exhibited duct-like structures (D and F, left panel), whereas in the normal pancreas, there was only weak Id-2 immunoreactivity in the ductal cells (E). Preabsorption with the Id-2-specific blocking peptide abolished the Id-2 immunoreactivity (F, right panel). Moderate to strong Id-3 immunoreactivity was present in the duct-like cancer cells (G and I, left panel). Moderate to strong Id-3 immunoreactivity was also present in the ductal cells of normal pancreatic tissue samples (H). Id-3 immunoreactivity was completely abolished by preabsorption with the Id-3 specific blocking peptide (I, right panel). A, D, and G constitute serial sections of a pancreatic cancer sample, revealing coexpression of the three Id proteins. Scale bars, 25 μm.

Cruz, CA) polyclonal antibodies were used for immunhistochemistry. These affinity-purified rabbit polyclonal antibodies specifically react with Id-1, Id-2, and Id-3, respectively, of human origin, as determined by Western blotting. Paraffin-embedded sections (4 μm) were subjected to immunostaining using the streptavidin-peroxidase technique. Where indicated, immunostaining for all three Id proteins was performed on serial sections. Endogenous peroxidase activity was blocked by incubation for 30 minutes with 0.3% hydrogen peroxide in methanol. Tissue sections were incubated for 15 minutes (23°C) with 10% normal goat serum and then incubated for 16 hours at 4°C with the indicated antibodies in PBS containing 1% bovine serum albumin. Bound antibodies were detected with biotinylated goat anti-rabbit IgG secondary antibodies and streptavidin-peroxidase complex, using diaminobenzidine tetrahydrochloride as the substrate. Sections were counterstained with Mayer's hematoxylin. Preabsorption with Id-1-, Id-2-, or Id-3-specific blocking peptides completely abolished immunoreactivity of the respective primary antibody. The immunohistochemical results were semiquantitatively analyzed as described previously.[29,30] The percentage of positive cancer cells was stratified into four groups: 0, no cancer cells exhibiting immunoreactivity; 1, <33% of the cancer cells exhibiting immunoreactivity; 2, 33 to 67% of the cancer cells exhibiting immunoreactivity; 3 >67% of the cancer cells exhibiting immunoreactivity. The intensity of the immunohistochemical signal was also stratified into four groups: 0, no immunoreactivity; 1, weak immunoreactivity; 2, moderate immunoreactivity; 3, strong immunoreactivity. Finally, the sum of the results of the cell score and the intensity score was calculated. Statistical analysis was performed with SigmaStat software. The rank sum test was used, and $P < 0.05$ was taken as the level of significance.

Results

Northern blot analysis of total RNA isolated from 12 normal pancreatic tissues and 16 pancreatic cancers revealed the presence of the 1.2-kb Id-1 transcript and the 1.6-kb Id2 mRNA transcript in 11 of the 12 normal pancreatic samples, and the 1.3-kb Id-3 mRNA transcript in all normal pancreatic samples (Figure 1A, 2). In the cancer tissues, Id-1 mRNA levels were elevated in 8 of 16 samples, Id-2 mRNA levels were elevated in 9 of these samples, and Id-3 mRNA levels were elevated in 6 of these samples (Figure 1A, 2). Concomitant overexpression of all three Id species was observed in 6 of the cancer samples (38%). In contrast, none of the Id mRNA species were overexpressed in CP by comparison with normal controls (Figure 1B, 2). Densitometric analysis of all of the autoradiograms indicated that there was a 6.5-fold increase ($P < 0.01$) in Id-1 mRNA levels, a fivefold increase ($P < 0.01$) in Id-2 mRNA levels, and a twofold increase ($P = 0.027$) in Id-3 mRNA levels in the pancreatic cancer samples in comparison to normal controls (Figure 2). In contrast, there was no statistically significant difference in the expression levels of Id-1, Id-2, and Id-3, in CP tissues in comparison to the corresponding levels in the normal pancreas (Figure 2).

Next, we assessed the expression of the three Id genes in 5 human pancreatic cancer cell lines by Northern and Western blot analyses. Id-1 mRNA was present at varying levels in all 5 cell lines (Figure 3). ASPC-1, CAPAN-1, MIA-PaCa-2, and PANC-1 expressed moderate to high levels of Id-1 mRNA, whereas COLO-357 cells

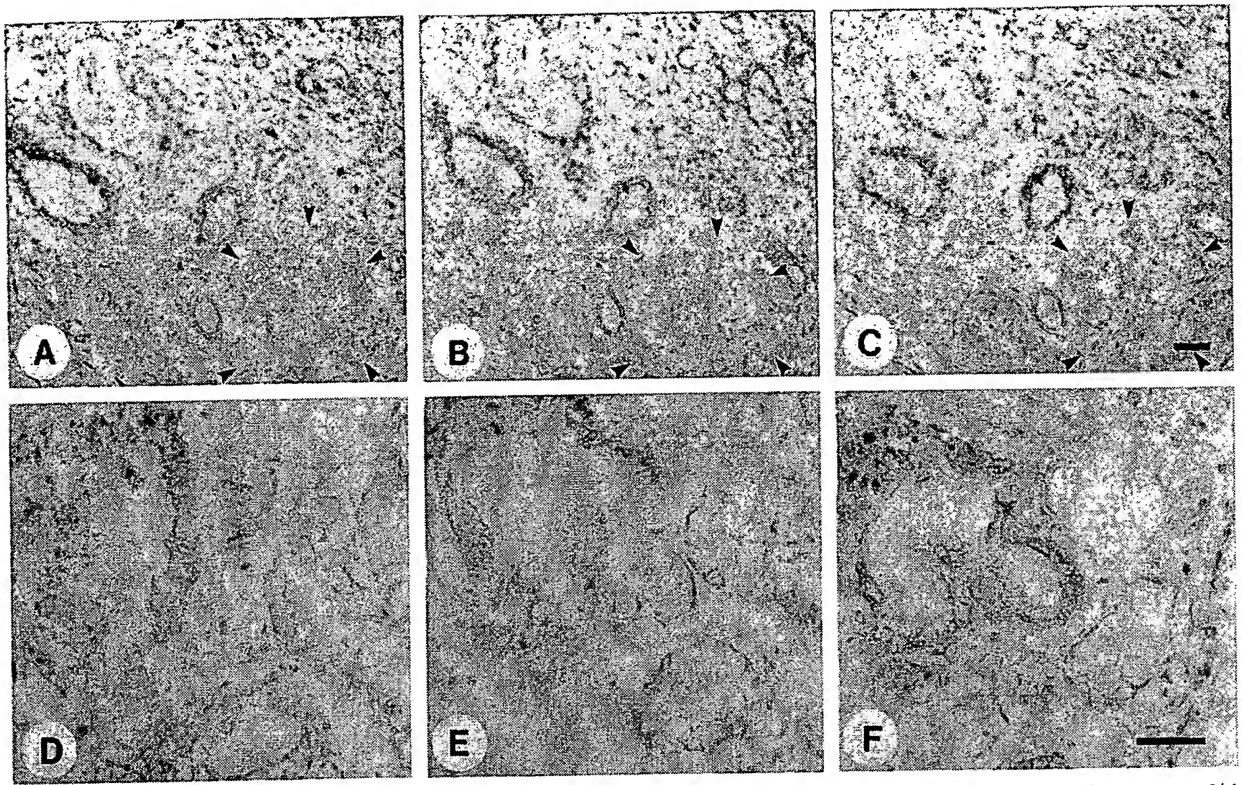


Figure 5. Immunohistochemistry of pancreatic cancer and dysplastic ducts in CP tissues. In the pancreatic cancer tissues (A-C) there was moderate to strong Id-1 (A), Id-2 (B), and Id-3 (C) immunoreactivity in the ductal cells in the areas adjacent to the cancer cells that exhibited CP-like alterations. Islet cells did not exhibit Id immunoreactivity (outlined by **solid arrowheads**). In the CP samples, moderate to strong Id-1 (D), Id-2 (E), and Id-3 (F) immunoreactivity was present in the cytoplasm of epithelial cells forming large dysplastic ducts. Scale bar, 25 μm.

expressed relatively low levels of this mRNA moiety. Western blotting with a highly specific anti-Id-1 antibody confirmed the presence of the approximately 14-kd Id-1 protein in the 4 cell lines that expressed high levels of Id-1 mRNA (Figure 3). Furthermore, the three cell lines with the highest Id-1 mRNA expression (CAPAN-1, MIA-PaCa-2, and PANC-1) also exhibited the highest Id-1 protein expression. Variable levels of the 1.6-kb Id-2 mRNA transcript were present in all 5 cell lines. In addition, a minor band of approximately 1.2 kb was visible in COLO-357 and MIA-PaCa-2 cells. Immunoblot analysis with a highly specific anti-Id-2 antibody revealed two bands of approximately 16 and 18 kd at relatively high levels in all of the cell lines with exception of PANC-1 cells, in which the 16-kd band was relatively faint (Figure 3). With the exception of MIA-PaCa-2 cells, there was a good correlation between Id-2 mRNA and protein levels (Figure 3). Id-3 mRNA was present at high levels in MIA-PaCa-2 cells, at moderate levels in COLO-357 cells, and at low levels in PANC-1 cells. Id-3 mRNA was not detectable in ASPC-1 and CAPAN-1 cells (Figure 3). Immunoblot analysis with a highly specific anti-Id-3 antibody revealed an approximately 14-kd band that was most abundant in MIA-PaCa-2 cells, and was also readily apparent in COLO-357 and PANC-1 cells. In contrast, only a faint Id-3 band was seen in ASPC-1 and CAPAN-1 cells. Thus, with the exception of PANC-1 cells, there was a good correlation between Id-3 mRNA and protein levels.

To determine the localization of Id-1, Id-2, and Id-3, immunostaining was carried out using the same highly specific anti-Id antibodies. In the pancreatic cancers, moderate to strong Id-1 immunoreactivity was present in the cancer cells in 9 of 10 randomly selected cancer samples. An example of moderate Id-1 immunoreactivity is shown in Figure 4A, and of strong immunoreactivity in Figure 4C (left panel). In contrast, in the normal pancreas, faint Id-1 immunoreactivity was present only in the ductal cells of pancreatic ducts (Figure 4B, arrowheads). Preabsorption with the Id-1-specific blocking peptide completely abolished the Id-1 immunoreactivity (Figure 4C, right panel). The cancer cells also exhibited strong Id-2 (Figure 4, D and F, left panel) and moderate to strong Id-3 immunoreactivity. An example of moderate Id-3 immunoreactivity is shown in Figure 4G, and of strong immunoreactivity in Figure 4I (left panel). In contrast, only faint Id-2 immunoreactivity was present in the ductal cells in the normal pancreas (Figure 4E), whereas Id-3 immunoreactivity in these cells was more variable and ranged from moderate to occasionally strong (Figure 4H). Islet cells and acinar cells were always devoid of Id immunoreactivity. Preabsorption of the respective antibody with the blocking peptides specific for Id-2 (Figure 4F, right panel) and Id-3 (Figure 4I, right panel) completely abolished immunoreactivity. Analysis of serial pancreatic cancer sections revealed that there was often colocalization of the

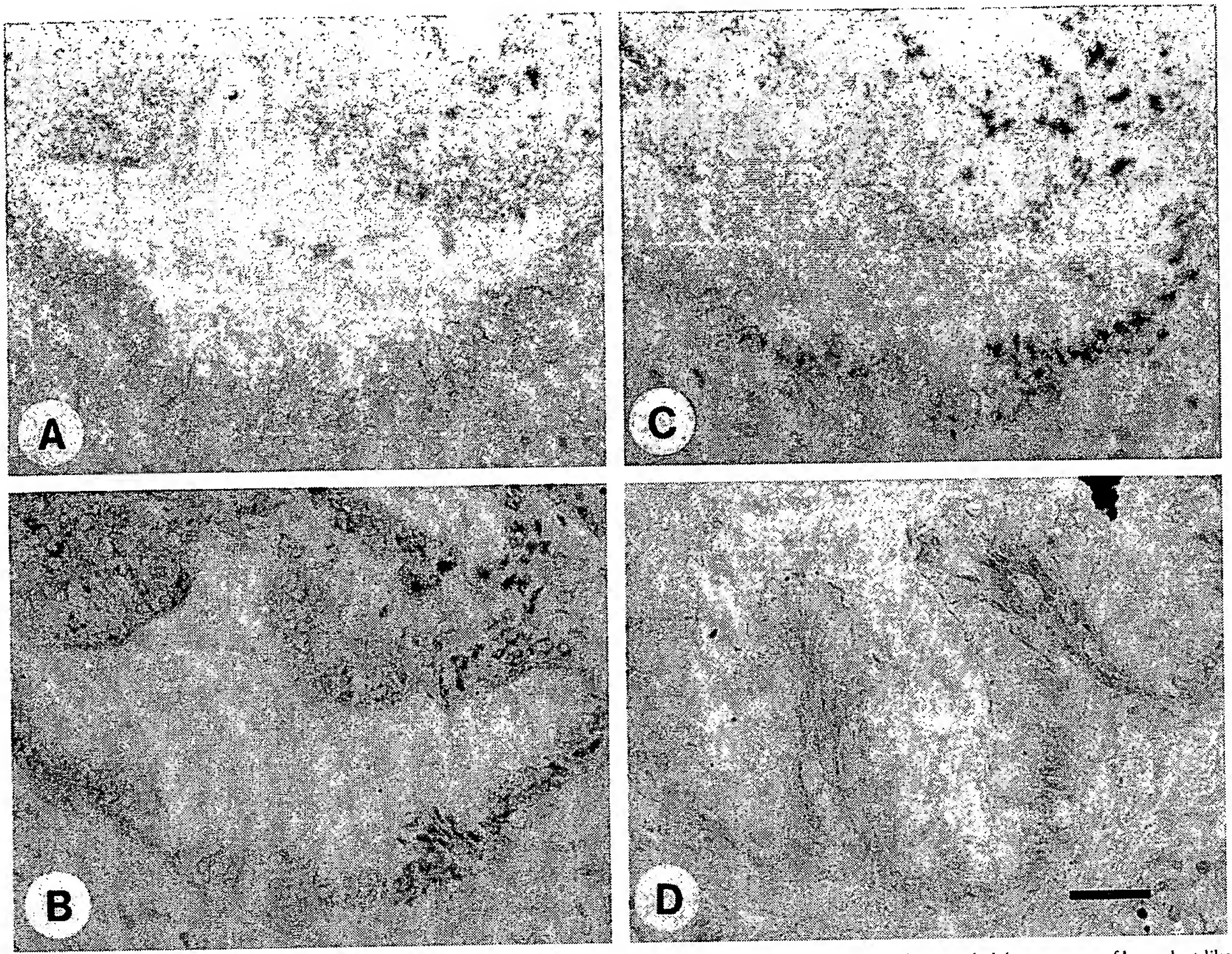


Figure 6. Immunohistochemistry of atypical papillary epithelium in CP tissues. Serial section analysis of some CP samples revealed the presence of large duct-like structures with atypical papillary epithelium. Mild to moderate Id-1 (A) and Id-2 (B) immunoreactivity and weak Id-3 (C) immunoreactivity was present in the cytoplasm of the cells forming these large ducts with papillary structures. Some CP samples also exhibited moderate Id-3 immunoreactivity in these cells (D). Scale bar, 25 μm.

three Id proteins. An example of serial sections from a pancreatic cancer tissue is shown in Figure 4, A, D, and G.

Id-1, Id-2, and Id-3 immunoreactivity was also present at moderate levels in the cytoplasm of ductal cells within CP-like areas adjacent to the cancer cells (Figure 5, A-C). As in the normal pancreas, islet cells (outlined by arrowheads) did not exhibit Id immunoreactivity. In 4 of 9 CP samples, there were foci of ductal cell dysplasia of relatively large interlobular ducts, all of which exhibited moderate to strong Id-1, Id-2, and Id-3 immunoreactivity (Figure 5, D-F). Five of 9 CP samples also contained foci of large ducts exhibiting atypical papillary epithelium. Serial section analysis of one of those CP samples revealed mild to moderate Id-1 and Id-2 immunoreactivity and weak Id-3 immunoreactivity in the cells of these atypical papillary ducts (Figure 6, A-C). In contrast, in some of these CP samples, moderate to strong Id-3 immunoreactivity was also observed (Figure 6D). However, most of the ductal cells forming the typical ductular structures of CP, such as large interlobular ducts and small proliferating ducts, exhibited generally only weak to occasionally moderate Id immunoreactivity (data not shown).

The immunohistochemical data for Id-1, Id-2, and Id-3 are summarized in Table 1. In the case of Id-1 and Id-2, the cancer cells as well as the dysplastic and atypical papillary ducts in CP exhibited a significantly higher score than the ductal cells in the normal pancreas. In contrast, due to the marked variability in Id-3 immunostaining in the normal pancreas, the differences between normal and cancer cells and normal and dysplastic cells did not achieve statistical significance.

Discussion

Id proteins constitute a family of HLH transcription factors that are important regulators of cellular differentiation and proliferation.[1,2] To date, four members of the human Id family have been identified.[1,10–12] Their expression is enhanced during cellular proliferation and in response to mitogenic stimuli,[19,31] and overexpression of Id genes inhibits differentiation and/or enhances proliferation in several different cell types.[15,32–34] The forced expression of Id-1 in mouse small intestinal epithelium results in

Table 1. Histological Scoring

		Id-1	Id-2	Id-3
Normal (*n* = 6)	Ductal cells	2.0 ± 0.4	2.3 ± 0.2	2.5 ± 0.9
Cancer (*n* = 10)	Cancer cells	4.5* ± 0.5	5.2§ ± 0.3	4.5 ± 0.6
CP (*n* = 9)	Typical CP lesions (*n* = 9)	2.7 ± 0.5	3.1 ± 0.6	3.4 ± 0.7
	Dysplastic ducts (*n* = 4)	5.3† ± 0.2	5.8‡ ± 0.2	5.3 ± 0.4
	Atypical papillary ducts (*n* = 5)	4.4‡ ± 0.2	5.2‡ ± 0.2	5.0 ± 0.4

Scoring of the histological specimens was performed as described in the Patients and Methods section. Values are the means ± SD of the number of samples indicated in parenthesis. *P* values are based on comparisons with the respective controls in the normal samples.
*, $P < 0.02$; $^{\dagger}P < 0.01$; $^{\ddagger}P = 0.004$; $^{\S}P = 0.001$.

adenoma formation in these animals.[35] The growth-promoting effects of Id genes are thought to occur through several mechanisms. For example, Id-2 can bind to members of the pRB tumor suppressor family, thus blocking their growth-suppressing activity,[20,21] and Id-1 and Id-2 can antagonize the bHLH-mediated activation of known inhibitors of cell cycle progression such as the cyclin-dependent kinase inhibitor p21.[23]

In the present study, we determined by Northern blot analysis that a significant percentage of human pancreatic cancers expressed increased Id-1, Id-2, and Id-3 mRNA levels. Increased expression was most evident for Id-1 (6.5-fold) and Id-2 (fivefold). In contrast, Id-3 mRNA levels were only twofold increased in the cancer samples, partly because this mRNA was present at relatively high levels in the normal pancreas. Immunhistochemical analysis confirmed the presence of Id-1, Id-2, and Id-3 in the cancer cells within the tumor mass, whereas in the normal pancreas faint Id-1 and Id-2 immunoreactivity and moderate to occasionally strong Id-3 immunoreactivity was present in some ductal cells. Pancreatic acinar and islet cells in the normal pancreas were devoid of Id-1, Id-2, and Id-3 immunoreactivity. In the cancer samples, all three Id proteins often colocalized in the cancer cells. Coexpression of all three Id genes was also observed in cultured pancreatic cancer cell lines, which often exhibited a close correlation between Id mRNA and protein expression. However, in MIA-PaCa-2 there was a divergence of Id-2 mRNA and protein levels, and in PANC-1 cells, Id-3 mRNA levels did not correlate well with Id-3 protein expression. These observations suggest that in these cells, the half-life of either Id mRNA or Id protein may be altered by comparison with the other cell lines. Interestingly, Id-2 immunoblotting revealed two closely spaced bands of approximately 16 and 18 kd in 4 of 5 cell lines. In view of the fact that two possible initiation codons have been reported for the Id-2 gene,[36] our observation raises the possibility that the two Id-2-immunoreactive bands may represent separate translation products of the Id-2 gene.

Pancreatic cancers often harbor p53 tumor suppressor gene mutations[37] and exhibit alterations in apoptosis pathways. Thus, these cancers often exhibit increased expression of anti-apoptotic proteins such as Bcl-2[38] and abnormal resistance to Fas-ligand-mediated apoptosis.[39] It has been shown recently that forced constitutive expression of Id genes together with the expression of anti-apoptotic genes such as Bcl-2 or BclX$_L$ can result in malignant transformation of human fibroblasts,[11] raising the possibility that the enhanced Id expression in pancreatic cancers together with increased expression of anti-apoptotic genes may contribute to the malignant potential of pancreatic cancer cells *in vivo*.

In the CP tissues there was no significant increase in Id-1, Id-2, and Id-3 mRNA levels in comparison to the normal pancreas. Immunohistochemical analysis of pancreatic cancer samples revealed colocalization of weak to moderate Id-1, Id-2, and Id-3 immunoreactivity in proliferating ductal cells in the CP-like regions adjacent to the cancer cells, indicating that Id expression was not restricted to the cancer cells. Similarly, analysis of CP samples indicated weak Id-1, Id-2, and Id-3 immunoreactivity in the cells of small proliferating ducts and large ducts without dysplastic changes. In general, there was a correlation between weak immunoreactivity and low Id mRNA levels. However, in samples that harbored large ducts with papillary structures there was moderate Id immunoreactivity, and in the cells forming dysplastic ducts there was moderate to strong Id immunoreactivity. In these CP samples, Id mRNA levels were relatively higher than in the CP samples that were devoid of these histological changes. Overall, however, increased Id expression, most notably of Id-1 and Id-2, distinguished a subgroup of pancreatic cancers from CP (Table 1).

Epidemiological studies have shown that the risk of developing pancreatic cancer is increased up to 16-fold in patients with pre-existing CP in comparison to the general population.[40] The mechanisms that contribute to neoplastic transformation in CP are not known. Although there is no established tumor progression model for pancreatic cancer, such as the adenoma-carcinoma sequence of colorectal carcinoma,[41] it is generally accepted that K-*ras* and p16 mutations occur relatively early in pancreatic carcinogenesis, whereas p53 mutations occur late in this process.[37,41–43] Increased Id expression may contribute to malignant transformation of cultured cell lines *in vitro*[11] and has been linked to cell invasion in a murine mammary epithelial cell line.[44] In view of the current findings that Id-1, Id-2, and Id-3 are overexpressed in pancreatic cancer and in dysplastic/metaplastic ducts in CP, these observations raise the possibility that elevated levels of Id-1, Id-2, and, to a lesser extent, Id-3 may represent relatively early markers of pancreatic malignant transformation and may contribute to the pathobiology of pancreatic cancer.

References

1. Jan YN, Jan LY: HLH proteins, fly neurogenesis, and vertebrate myogenesis. Cell 1993, 75:827–830
2. Olson EN, Klein WH: bHLH factors in muscle development: dead lines and commitments, what to leave in and what to leave out. Genes Dev 1994, 8:1–8
3. Begley CG, Aplan PD, Denning SM, Haynes BF, Waldmann TA, Kirsch IR: The gene SCL is expressed during early hematopoiesis and encodes a differentiation-related DNA-binding motif. Proc Natl Acad Sci USA 1989, 86:10128–10132
4. Johnson JE, Birren SJ, Anderson DJ: Two rat homologues of Drosophila achaete-scute specifically expressed in neuronal precursors. Nature 1990, 346:858–861
5. Weintraub H: The MyoD family and myogenesis: redundancy, networks, and thresholds. Cell 1993, 75:1241–1244
6. Hu JS, Olson EN, Kingston RE: HEB, a helix-loop-helix protein related to E2A, and ITF2 that can modulate the DNA-binding ability of myogenic regulatory factors. Mol Cell Biol 1992, 12:1031–1042
7. Langlands K, Yin X, Anand G, Prochownik EV: Differential interactions of Id proteins with basic-helix-loop-helix transcription factors. J Biol Chem 1997, 272:19785–19793
8. Murre C, Bain G, van Dijk MA, Engel I, Furnari BA, Massari ME, Matthews JR, Quong MW, Rivera RR, Stuiver MH: Structure and function of helix-loop-helix proteins. Biochim Biophys Acta 1994, 1218:129–135
9. Murre C, McCaw PS, Vaessin H, Caudy M, Jan LY, Jan YN, Cabrera CV, Buskin JN, Hauschka SD, Lassar AB, Baltimore D: Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. Cell 1989, 58: 537–544
10. Benezra R, Davis RL, Lockshon D, Turner DL, Weintraub H: The protein Id: a negative regulator of helix-loop-helix DNA binding proteins. Cell 1990, 61:49–59
11. Norton JD, Atherton GT: Coupling of cell growth control and apoptosis functions of Id proteins. Mol Cell Biol 1998, 18:2371–2381
12. Norton JD, Deed RW, Craggs G, Sablitzky F: Id helix-loop-helix proteins in cell growth and differentiation. Trends Cell Biol 1998, 8:58–65
13. Christy BA, Sanders LK, Lau LF, Copeland NG, Jenkins NA, Nathans D: An Id-related helix-loop-helix protein encoded by a growth factor-inducible gene. Proc Natl Acad Sci USA 1991, 88:1815–1819
14. Kawaguchi N, DeLuca HF, Noda M: Id gene expression and its suppression by 1,25-dihydroxyvitamin D3 in rat osteoblastic osteosarcoma cells. Proc Natl Acad Sci USA 1992, 89:4569–4572
15. Kreider BL, Benezra R, Rovera G, Kadesch T: Inhibition of myeloid differentiation by the helix-loop-helix protein Id. Science 1992, 255: 1700–1702
16. Le Jossic C, Ilyin GP, Loyer P, Glaise D, Cariou S, Guguen-Guillouzo C: Expression of helix-loop-helix factor Id-1 is dependent on the hepatocyte proliferation and differentiation status in rat liver and in primary culture. Cancer Res 1994, 54:6065–6068
17. Sun XH, Copeland NG, Jenkins NA, Baltimore D: Id proteins Id1 and Id2 selectively inhibit DNA binding by one class of helix-loop-helix proteins. Mol Cell Biol 1991, 11:5603–5611
18. Wilson RB, Kiledjian M, Shen CP, Benezra R, Zwollo P, Dymecki SM, Desiderio SV, Kadesch T: Repression of immunoglobulin enhancers by the helix-loop-helix protein Id: implications for B-lymphoid-cell development. Mol Cell Biol 1991, 11:6185–6191
19. Hara E, Yamaguchi T, Nojima H, Ide T, Campisi J, Okayama H, Oda K: Id-related genes encoding helix-loop-helix proteins are required for G1 progression and are repressed in senescent human fibroblasts. J Biol Chem 1994, 269:2139–2145
20. Iavarone A, Garg P, Lasorella A, Hsu J, Israel MA: The helix-loop-helix protein Id-2 enhances cell proliferation and binds to the retinoblastoma protein. Genes Dev 1994, 8:1270–1284
21. Lasorella A, Iavarone A, Israel MA: Id2 specifically alters regulation of the cell cycle by tumor suppressor proteins. Mol Cell Biol 1996, 16:2570–2578
22. Peverali FA, Ramqvist T, Saffrich R, Pepperkok R, Barone MV, Philipson L: Regulation of G1 progression by E2A and Id helix-loop-helix proteins. EMBO J 1994, 13:4291–4301
23. Prabhu S, Ignatova A, Park ST, Sun XH: Regulation of the expression of cyclin-dependent kinase inhibitor p21 by E2A and Id proteins. Mol Cell Biol 1997, 17:5888–5896
24. Warshaw AL, Fernandez-del Castillo C: Pancreatic carcinoma. N Engl J Med 1992, 326:455–465
25. Korc M: Role of growth factors in pancreatic cancer. Surg Oncol Clin North Am 1998, 7:25–41
26. Kleeff J, Ishiwata T, Friess H, Büchler MW, Israel MA, Korc M: The helix-loop-helix protein Id2 is overexpressed in human pancreatic cancer. Cancer Res 1998, 58:3769–3772
27. Oertel JE, Heffes CS, Oertel YC: Pancreas. Diagnostic Surgical Pathology. Edited by SS Sternberg. New York, Raven Press, 1989, pp 1057–1093
28. Korc M, Chandrasekar B, Yamanaka Y, Friess H, Büchler MW, Beger HG: Overexpression of the epidermal growth factor receptor in human pancreatic cancer is associated with concomitant increase in the levels of epidermal growth factor and transforming growth factor α. J Clin Invest 1992, 90:1352–1360
29. Saeki T, Stromberg K, Qi CF, Gullick WJ, Tahara E, Normanno N, Ciardiello F, Kenney N, Johmson GR, Salomon DS: Differential immunohistochemical detection of amphiregulin and cripto in human normal colon and colorectal tumors. Cancer Res 1992, 52:3467–3473
30. Cantero D, Friess H, Deflorin J, Zimmermann A, Bründler MA, Riesle E, Korc M, Büchler MW: Enhanced expression of urokinase plasminogen activator and its receptor in pancreatic carcinoma. Br J Cancer 1997, 75:388–395
31. Desprez PY, Hara E, Bissell MJ, Campisi J: Suppression of mammary epithelial cell differentiation by the helix-loop-helix protein Id-1. Mol Cell Biol 1995, 15:3398–3404
32. Shoji W, Yamamoto T, Obinata M: The helix-loop-helix protein Id inhibits differentiation of murine erythroleukemia cells. J Biol Chem 1994, 269:5078–5084
33. Cross JC, Flannery ML, Blanar MA, Steingrimsson E, Jenkins NA, Copeland NG, Rutter WJ, Werb Z: Hxt encodes a basic helix-loop-helix transcription factor that regulates trophoblast cell development. Development 1995, 121:2513–2523
34. Sun XH: Constitutive expression of the Id1 gene impairs mouse B cell development. Cell 1994, 79:893–900
35. Wice BM, Gordon JI: Forced expression of Id-1 in the adult mouse small intestinal epithelium is associated with development of adenomas. J Biol Chem 1998, 273:25310–25319
36. Barone MV, Pepperkok R, Peverali FA, Philipson L: Id proteins control growth induction in mammalian cells. Proc Natl Acad Sci USA 1994, 91:4985–4988
37. Barton CM, Staddon SL, Hughes CM, Hall, PA, O'Sullivan C, Kloppel G, Theis B, Russell RC, Neoptolmos J, Williamson RCN, Lane DP, Lemoine NR: Abnormalities of the p53 tumour suppressor gene in human pancreatic cancer. Br J Cancer 1991, 64:1076–1082
38. Ohshio G, Suwa H, Imamura T, Yamaki K, Tanaka T, Hashimoto Y, Imamura M: An immunohistochemical study of bcl-2 and p53 protein expression in pancreatic carcinomas. Scand J Gastroenterol 1998, 33:535–539
39. Ungefroren H, Voss M, Jansen M, Roeder C, Henne-Bruns D, Kremer B, Kalthoff H: Human pancreatic adenocarcinomas express Fas and Fas ligand yet are resistant to Fas-mediated apoptosis. Cancer Res 1998, 58:1741–1749
40. Niederau C, Niederau MC, Heintges T, Lüthen R: Epidemiology: relation between chronic pancreaitis and pancreatic carcinoma. Cancer of the Pancreas. Edited by HG Beger, MW Büchler, MH Schoenberg. Ulm, Germany, Universitätsverlag Ulm GmbH, 1996, pp 6–9
41. Moskaluk CA, Kern SE: Molecular genetics of pancreatic carcinoma. Pancreatic Cancer: Pathogenesis, Diagnosis, and Treatment. Edited by HA Reber. Totowa, NJ, Humana Press, 1998, pp 3–20
42. Moskaluk CA, Hruban RH, Kern SE: p16 and K-ras gene mutations in the intraductal precursors of human pancreatic adenocarcinoma. Cancer Res 1997, 57:2140–2143
43. Tada M, Ohashi M, Shiratori Y, Okudaira T, Komatsu Y, Kawabe T, Yoshida H, Machinami R, Kishi K, Omata M: Analysis of K-ras gene mutation in hyperplastic duct cells of the pancreas without pancreatic disease. Gastroenterology 1996, 110:227–231
44. Desprez PY, Lin CQ, Thomasset N, Sympson CJ, Bissell MJ, Campisi J: A novel pathway for mammary epithelial cell invasion induced by the helix-loop-helix protein Id-1. Mol Cell Biol 1998, 18:4577–4588

American Journal of Pathology, Vol. 161, No. 4, October 2002
Copyright © American Society for Investigative Pathology

Distinct and Complementary Information Provided by Use of Tissue and DNA Microarrays in the Study of Breast Tumor Markers

Christophe Ginestier,[*]
Emmanuelle Charafe-Jauffret,[*†]
François Bertucci,[*‡] François Eisinger,[§]
Jeannine Geneix,[*] Didier Bechlian,[*]
Nathalie Conte,[*] José Adélaïde,[*] Yves Toiron,[*†]
Catherine Nguyen,[¶] Patrice Viens,[‡]
Marie-Joelle Mozziconacci,[*†] Rémi Houlgatte,[¶]
Daniel Birnbaum,[*] and Jocelyne Jacquemier[*†]

From the Département d'Oncologie Moléculaire,[*] Institut Paoli-Calmettes and Institut National de la Santé et de la Recherche Medical U119, IFR57, Marseille; the Departements de Biopathologie[†] and de Dépistage et Prévention,[§] Institut Paoli-Calmettes, Marseille; the Département d'Oncologie Médicale,[‡] Institut Paoli-Calmettes, Université de la Méditerranée, Marseille; and the Laboratoire Technologies Avancées pour le Génome et la Clinique,[¶] Centre d'Immunologie de Marseille-Luminy, Luminy, Marseille, France

Emerging high-throughput screening technologies are rapidly providing opportunities to identify new diagnostic and prognostic markers and new therapeutic targets in human cancer. Currently, cDNA arrays allow the quantitative measurement of thousands of mRNA expression levels simultaneously. Validation of this tool in hospital settings can be done on large series of archival paraffin-embedded tumor samples using the new technique of tissue microarray. On a series of 55 clinically and pathologically homogeneous breast tumors, we compared for 15 molecules with a proven or suspected role in breast cancer, the mRNA expression levels measured by cDNA array analysis with protein expression levels obtained using tumor tissue microarrays. The validity of cDNA array and tissue microarray data were first verified by comparison with quantitative reverse transcriptase-polymerase chain reaction measurements and immunohistochemistry on full tissue sections, respectively. We found a good correlation between cDNA and tissue array analyses in one-third of the 15 molecules, and no correlation in the remaining two-thirds. Furthermore, protein but not RNA levels may have prognostic value; this was the case for MUC1 protein, which was studied further using a tissue microarray containing ~600 tumor samples. For THBS1 had prognostic value. Thus, differences extended to clinical prognostic information obtained by the two methods underlining their complementarity and the need for a global molecular analysis of tumors at both the RNA and protein levels. ***(Am J Pathol 2002, 161:1223–1233)***

The development of genomic, technological, and bioinformatic tools have allowed progress in cancer research. DNA arrays are currently the most used of the new high-throughput methods to analyze the molecular complexity of tumors. Several studies have showed their potential in many types of human cancers.[1–4] Even if the clinical benefits for patients remain to be demonstrated, the first results are very encouraging. DNA arrays-based gene expression profiles are improving our understanding of the disease as well as tumor taxonomy by identifying new diagnostic or prognostic subclasses unrecognized by usual parameters. They are expected to lead to the discovery of new potential therapeutic targets, to accurate predictions of survival and response to a given treatment, and eventually to the delivery of a therapy appropriate to each individual patient.

Once a potential marker is identified by this technique, an important next step is its validation and introduction in routine tests in hospital settings.[5,6] There, cDNA arrays are not the method of choice because they are still expensive, time-consuming, complex, and require frozen material not always available. Validation studies have been done traditionally by immunohistochemistry (IHC) on paraffin-embedded tissues allowing analysis of many archived samples with a long follow-up. Until recently, pathologists examined sections of tumor slide by slide. Today, the recently developed tissue microarray (TMA) technology[7–9] allows the simultaneous analysis of thousands of tumor samples arrayed onto glass slides. This may facilitate the search for correlations between

Supported by INSERM, Institut Paoli-Calmettes, and grants from l'Association pour la Recherche sur le Cancer and la Ligue Nationale contre le Cancer.

ECJ and FB contributed equally to this work.

Accepted for publication June 20, 2002.

Address reprint requests to Daniel Birnbaum, U119 INSERM, Institut

molecular alterations and the histoclinical features of the tumors.

In a recent cDNA array-based, prognosis-oriented study of 55 localized breast carcinoma samples,[10] we identified two clusters of discriminator genes (named I and II) the differential expression of which allowed to distinguish subclasses of tumors with significantly different clinical outcome after adjuvant chemotherapy. The aim of the present study was to validate some of these data using TMAs and to evaluate the interest and limitations of this technology as a validation tool. Cylinders from the same 55 tumors were arrayed in a specific tissue-microarray and studied by IHC using antibodies directed against proteins encoded by some of our discriminator genes.

Materials and Methods

Mammary Carcinoma Cell Lines

Nine established mammary carcinoma cell lines were used as positive controls for expression of various genes or proteins. They included: BT-474, MCF-7, MCF-10F, MDA-MB-157, MDA-MB-175, MDA-MB-231, MDA-MB-453, BrCa-MZ-02,[11] and HBL-100. All cell lines are derived from carcinomas except HBL-100 and MCF-10F. They were obtained from the American Type Culture Collection, Rockville, MD (*http://www.atcc.org/*) and grown using the recommended culture conditions.

Breast Tumor Samples and Characteristics of Patients

Tumor samples were obtained from 55 women treated at the Institut Paoli-Calmettes. Inclusion criteria were: 1) localized breast cancer treated with adjuvant anthracyclin-based chemotherapy in addition to loco-regional treatment; 2) tumor material quickly macrodissected and frozen in liquid nitrogen and stored at −160°C; and 3) patient follow-up of 48 months or more after diagnosis. In addition to the axillary lymph node status, four poor prognosis criteria were used to determine whether adjuvant chemotherapy should be administered: patient age less than 40 years, pathological tumor size greater than 20 mm, Scarff-Bloom-Richardson grade equal to 3, and negative estrogen receptor (ER) status as evaluated by IHC with a positivity cutoff value of 1%. Women who received chemotherapy were those with either node-positive tumors or node-negative tumors and one of the poor prognosis criteria if nonmenopausal or two criteria if menopausal. All tumor sections were *de novo* reviewed by a pathologist (JJ) before analysis; all samples contained more than 50% tumor cells. Tumors were infiltrating adenocarcinomas including, according to the World Health Organization histological typing, 42 ductal, 5 lobular, 5 mixed, and 3 medullary carcinomas.

A second series of breast tumors was analyzed. It was constituted by 592 localized forms of breast cancer col-

nitrogen (the 55 tumors previously described were included in this array). There were 401 ductal, 77 lobular, 40 mixed, 4 medullary carcinomas, and 70 other histological types. A total of 297 tumors were node positive and 450 were positive for ER.

Extraction of RNA from Frozen Tissue

Total RNA was extracted from tumor samples by standard methods, as previously described.[12] RNA integrity was controlled on denaturing formaldehyde-agarose gel electrophoresis and Northern blots using a 28S-specific oligonucleotide.

DNA Arrays

DNA arrays were made in our facility (Technologies Avanceés pour le Génome et la Clinique)). Nylon filter preparation with spotted polymerase chain reaction (PCR) products derived from ~1000 selected candidate cancer genes, ^{33}P radioactive hybridization, and data acquisition, normalization, and analysis have been described elsewhere[13,14] and can also be consulted on our web site (*http:/tagc.univ-mrs.fr/pub/Cancer/*).

Reverse Transcription

RNA extracted from frozen tissue was reverse-transcribed in a final volume of 20 μl containing 1× reverse transcriptase (RT)-PCR buffer (Invitrogen Corp., Carlsbad, CA) , 5 mmol/L $MgCl_2$ (Invitrogen), 1 mmol/L dXTP (Roche Diagnostics, Meylan, France), 10 mmol/L dithiothreitol (Invitrogen), 5 μmol/L random hexamers (Roche), 20 U of RNase inhibitor (Promega Biosciences, Madison, WI) , 200 U of superscript reverse transcriptase (Invitrogen), and 1 μg of total RNA (calibration curve points and patient samples). Samples were incubated at 20°C for 10 minutes and 42°C for 45 minutes; reverse transcriptase was inactivated by heating at 99°C for 3 minutes and cooling at 4°C for 5 minutes.

Real-Time Quantitative RT-PCR (RQ-PCR)

RQ-PCR analyses for *ERBB2*, *MUC1*, and *TBP* (TATA box binding protein) mRNA were done using the ABI PRISM 7700 Sequence Detection System instrument and software (Perkin Elmer Applied Biosystems, Foster City, CA). Conditions for the analysis of these markers have been described.[15,16] Primers and probes for the TaqMan system were designed to meet specific criteria by using Primer Express software (Perkin Elmer) and were synthesized by Genset (Genset Olijos, La Jolla, CA, USA) for the primers and by Roche for the probes. The 5′- and 3′-end nucleotides of the probe were labeled with a reporter (FAM, 6-carboxy-fluorescein) and a quencher dye (TAMRA, 6-carboxy-tetramethylrhodamine). The sequences of the PCR primer pairs and fluorogenic probes used for each gene

Table 1. Sequences of Oligonucleotide Primers and Probes Used in RQ-PCR Experiments

Gene	Oligonucleotide	Sequence	PCR product size
ERBB2	Forward primer	5′-AGCCGCGAGCACCCAAGT-3′ (exon 1)	147 bp
	Reverse primer	5′-TTGGTGGGCAGGTAGGTGAGTT-3′ (exon 2)	
	Probe	5′-CCTGCCAGTCCCGAGACCCACCT-3′	
MUC1	Forward primer	5′-ACCATCCTATGAGCGAGTACC-3′ (exon 6)	107 bp
	Reverse primer	5′-GTTTCTGCAGGTAATGGTGGC-3′ (exon 7)	
	Probe	5′-CCCATGGGCGCTATGTGCCC-3′	
TBP	Forward primer	5′-CACGAACCACGGCACTGATT-3′	89 bp
	Reverse primer	5′-TTTTCTTGCTGCCAGTCTGGAC-3′	
	Probe	5′-TGTGCACAGGAGCCAAGAGTGAAG-3′	

Bank accession no. M11730, *MUC1* GenBank accession no. JO5581, *TBP* GenBank accession no. X54993. The precise amount of total RNA added to each reaction mix (based on absorbance) and its quality (ie, lack of extensive degradation) are both generally difficult to assess. Therefore, the relative expression level of the gene of interest was computed with respect to the internal standard *TBP* to normalize for variations in the quality of RNA and the amount of input cDNA. Ct (threshold cycle) was used for quantification of the input target number and all experiments were done with duplicates for each data point. All patient samples with a variation >1 Ct for the duplicate were retested. For each experimental sample, the amount of target and endogenous reference was determined from a standard curve. The standard curve was constructed with fivefold serial dilutions of cDNA (1000 ng to 1 ng) from BT-474 (for *ERBB2*) and MCF-7 (for *MUC1*) breast carcinoma cell lines, respectively. The relative target gene expression in a tested sample was normalized using a calibrator sample, ie, the HME1 human primary mammary epithelial cell line (Clontech). The level of expression of the target gene was given by the N-ratio, in which each normalized gene value (*ERBB2*, *MUC1*) was divided by a calibrator normalized gene value (*TBP*).

$$N_{ERBB2} = \frac{ERBB2_{SAMPLE}}{TBP_{SAMPLE}} \bigg/ \frac{ERBB2_{CALIBRATOR}}{TBP_{CALIBRATOR}}$$

$$N_{MUC1} = \frac{MUC1_{SAMPLE}}{TBP_{SAMPLE}} \bigg/ \frac{MUC1_{CALIBRATOR}}{TBP_{CALIBRATOR}}$$

PCR was done with 1× TaqMan Universal PCR Master Mix (Perkin Elmer), 300 nmol/L of primers, 200 nmol/L of the probe, and 1 μl of each appropriately diluted reverse transcription sample in a 25-μl final reaction mixture. After a 2-minute incubation at 50°C to allow for uracyl N-glycosylate cleavage, AmpliTaq Gold was activated by an incubation for 10 minutes at 95°C. Each of the 40 PCR cycles consisted of 15 seconds of denaturation at 95°C and hybridization of probe and primers for 1 minute at 60°C.

TMA Construction

TMAs were prepared as described[9] with slight modifications. For each tumor, three representative tumor areas were carefully selected from a hematoxylin- and eosin-stained section of a donor block. Core cylinders with a diameter of 0.6 mm each were punched from each of these areas and deposited into a recipient paraffin block

Table 2. List of Proteins Tested by Immunohistochemistry and Characteristics of the Corresponding Antibodies

Protein	Antibody	Origin	Clone	Dilution
Angiogenin (ANG)	Rabbit polyclonal	Santa Cruz Biotechnology	sc-9044	1/20
BCL2	mmab	DAKO	124	1/100
E Cadherin (CDH1)	mmab	Transduction Laboratories	36	1/2000
ERBB2	mmab	Novocastra Laboratories Ltd.	CB 11	1/500
ERBB2	mmab	Oncogene Research Products	3B5	1/500
ERBB2	Rabbit polyclonal	DAKO	AO 485	1/1000
Estrogen receptor (ESR1/ER)	mmab	Novocastra Laboratories Ltd.	6F11	1/60
FGFR1	Rabbit polyclonal	Santa Cruz Biotechnology	sc-121	1/200
GATA3	mmab	Santa Cruz Biotechnology	sc-268	1/100
Ki67	mmab	DAKO	KI-67	1/100
Melan A/MART1 (MLANA)	mmab	DAKO	A103	1/2
MUC1	mmab	Transgen	H23	1/1000
P53	mmab	Immunotech	DO-1	1/4
Progesterone receptor (PR)	mmab	DAKO	PgR 636	1/80
Prolactin receptor (PRLR)	mmab	NeoMarkers	B6.2	1/200
Transforming acidic coiled-coil 1 TACC1	Rabbit polyclonal	Upstate Biotechnology	07-229	1/200
Transforming acidic coiled-coil 2 TACC2	Rabbit polyclonal	Upstate Biotechnology	07-228	1/40
Thrombospondin 1 (THBS1)	mmab	Oncogene Research Products	46.4	1/10

using a specific arraying device (Beecher Instruments, Silver Spring, MD). In addition to tumor tissues, the recipient block also received normal breast tissue and cell line pellets. Five-μm sections of the resulting microarray block were made and used for IHC analysis after transfer to glass slides. Two TMAs were prepared; the first one contained the 55 tumors studied by cDNA arrays (with three cores per sample) and controls, the second one was used for MUC1 study and contained 592 tumor samples (with one core per sample) and controls.

Antibodies and IHC

The characteristics of the antibodies used are listed in Table 2. IHC was performed on 5-μm sections of formalin-embedded tissue specimens. They were deparaffinized in histolemon (Carlo Erba Reagenti, Rodano, Italy) and rehydrated in graded alcohol. Antigen enhancement was done by incubating the sections in target retrieval solution (DAKO, Copenhagen, Denmark) as recommended except for prolactin receptor, in which pretreatment was done with incubation in pepsin (Zymed Laboratories, South San Francisco, CA), for 30 minutes at 37°C, and for MUC1, in which no pretreatment was done. Slides were then transferred to a DAKO autostainer. Staining was done at room temperature as follows: after washes in phosphate buffer, followed by quenching of endogenous peroxidase activity by treatment with 0.1% H_2O_2, slides were first incubated with blocking serum (DAKO) for 10 minutes and then with the affinity-purified antibody for 1 hour. After washes, slides were incubated with biotinylated antibody against rabbit Ig for 20 minutes followed by streptavidin-conjugated peroxidase (DAKO LSAB^{R}2 kit). Diaminobenzidine or 3-amino-9-ethylcarbazole was used as the chromogen, counterstained with hematoxylin, and coverslipped using Aquatex (Merck, Darmstadt, Germany) mounting solution. Slides were evaluated under a light microscope by two pathologists (EC-J, JJ).

Immunoreactivities were classified by estimating the percentage (P) of tumor cells showing characteristic staining (from undetectable level or 0%, to homogeneous staining or 100%) and by estimating the intensity (I) of staining (1, weak staining; 2, moderate staining; or 3, strong staining. The cutoff values were the same for all markers tested. Results were scored by multiplying the percentage of positive cells by the intensity, ie, by the so-called quick score (Q) ($Q = P \times I$; maximum = 300). For Ki67, only the percentage (P) of tumor cells was estimated, because intensity does not vary. Expression levels allowed to group tumors into four categories: negative expression ($Q = 0$ or $P = 0$ for Ki67), weak expression ($0 < Q \leq 120$ or $0 < P < 25$ for Ki67), moderate expression ($120 < Q \leq 210$ or $25 \leq P < 60$ for Ki67) and strong expression ($210 < Q \leq 300$ or $60 \leq P \leq 100$ for Ki67). Because of its prognostic impact the topographical localization of MUC1 was taken into account and expressed in four categories: absence, apical, circumferential membrane, and cytoplasmic

IHC on Full Tissue Sections

To validate the use of TMAs for immunophenotyping, we compared the protein expression levels of ER, progesterone receptor, P53, and BCL2, on full tissue sections and on TMAs for the group of 55 tumors. The data on full sections were compared to the mean of intensities of the three 0.6-mm core biopsies for 47 cases, or of only two core biopsies for 8 cases.

Statistical Analysis

The concordance between RNA expression levels measured by real-time quantitative RT-PCR and cDNA arrays was examined using Spearman's rank correlation. Comparison between IHC data from full sections and TMAs analyses was measured using κ statistics (a κ value >0.7 indicated a strong association). Contingency table analysis was used to analyze the relationship between protein expression obtained by IHC on TMAs and RNA expression obtained with cDNA arrays (total chi-square test). Survival analysis used the Kaplan-Meier method and survival curves were compared using the log-rank test (a P value <0.5 was considered as significant). All P values were two-sided. To assess the relationship between two variables assumed to be related (ie, the co-regulated molecules), simple linear regression analyses were performed using Excel Software (Microsoft). For these tests, (O,O) points were removed; the relationship tested was thus for cases with at least one positive value. Each result is given with: N the sample size, a the slope of the regression line, the P value and $r2$, the coefficient of determination. Thus, for each positive comparison a linear relationship can be determined (eg, $y = 0.8x + 20$ means BCL2 = 0.8ER + 20).

Results

Selection of Molecules

We previously analyzed the mRNA expression profiles of ~1000 selected genes in 55 breast carcinoma samples using home-made cDNA arrays. Tumors were homogeneous with respect to histological and clinical parameters, and all patients had received adjuvant anthracyclin-based chemotherapy. Detailed results are described elsewhere.[10] Briefly, molecular profiling combined with hierarchical clustering allowed the identification, among this set of poor-prognosis localized breast cancers, of new subclasses distinct with respect to overall and metastasis-free survivals. Such a classification resulted from the differential expression of two discriminator gene clusters (named I and II) and was not possible using classical prognostic factors of disease. Cluster I included the *ESR1* gene encoding ER-α. For the present study, we selected 10 of these genes. Interestingly, six of them (*BCL2*, *ERBB2*, *ESR1*, *GATA3*, *MUC1*, *PRLR*) have also been frequently identified as discriminator genes in expression-profiling studies of breast cancer that have addressed the prognosis issue.[4,17–19] These genes were

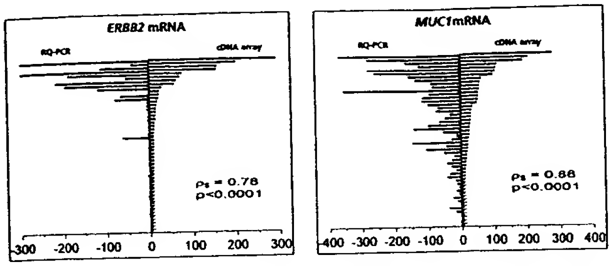


Figure 1. Expression levels of *ERBB2* and *MUC1* mRNA levels measured by cDNA array analysis and real-time quantitative PCR amplification. *ERBB2* and *MUC1* mRNA expression levels measured using cDNA arrays (artificially ×30 for visual effect) (**left**) and real-time quantitative PCR amplification (artificially ×30 for visual effect) (**right**). Results for each tumor (from top to bottom) are represented as **opposite bars**. For *ERBB2*: $\rho_s = 0.78$, $P < 0.0001$; for *MUC1*: $\rho_s = 0.88$, $P < 0.0001$.

thus interesting candidates for further investigation. In addition, other molecules, such as CDH1, Ki67, TP53, progesterone receptor, TACC1,[20] and TACC2, were retained because of a known or suspected role in breast cancer. The selection criteria for all molecules also included availability of a commercial antibody. The complete list of the corresponding proteins tested in the following experiments is given in Table 2.

Validation of cDNA Array Data with RQ-PCR

Our cDNA array analyses regularly included extensive experiments and controls designed to ensure reproducibility and reliability of expression measurements.[1,13,14,21] Nevertheless, we sought to further validate our data by comparing RNA expression levels of two genes, *ERBB2* and *MUC1*, as measured by cDNA array, to those obtained by RQ-PCR.

RNA from 50 of 55 samples (RNA was no longer available for five cases) was reverse-transcribed and PCR amplification of *ERBB2* and *MUC1* cDNA was done using a TaqMan device. For *ERBB2*, 41 tumors displayed mRNA expression levels comparable to normal breast and HME1 control cell line, whereas nine samples (18%) showed overexpression. For *MUC1*, 17 tumors displayed mRNA expression levels comparable to normal breast and HME1 control cell line, whereas 33 samples (66%) showed overexpression. As shown in Figure 1, mRNA expression levels obtained with both methods were highly similar (Spearman test: *ERBB2*, $\rho_s = 0.78$, $P < 0.0001$; *MUC1*, $\rho_s = 0.88$, $P < 0.0001$), further suggesting reliability of our cDNA array data.

TMA Analysis and Validation of Data

To validate our TMA analyses, we compared the expression of four selected proteins (BCL2, ER, P53, progesterone receptor) measured by IHC using either standard full tissue sections or TMAs in the panel of 55 breast tumors. For BCL2 expression, 38 cases (69%) showed positive cytoplasm staining, whereas 17 cases (31%) were negative on analysis of full sections. In comparison, 37 cases (67%) were positive and 18 cases were negative (33%) on TMA. Overall, the concordance was 91% and the nonconcordance was 9% (five cases), resulting in a strong statistical association between the two methods (κ value, 0.78). An even better correlation was found for nuclear expressions of ER, P53, and progesterone receptor, with only 3 discordant cases of 55 for each of them (concordance, 95%; Kappa values, 0.86 to 0.88). This high degree of concordance between IHC on full sections and on TMAs justified further use of TMAs.

Analysis of Breast Tumors Using TMAs

Fifteen proteins, including the four previously cited, were tested by IHC on TMAs. Most of them corresponded to genes we had identified in our two discriminator gene clusters I and II.[10] Other tested molecules corresponded to proteins of interest in breast cancer. Immunostainings were evaluated by the quick score (except for Ki67). Results are shown in Table 3 and Figure 2.

Table 3. Results of IHC Stainings on Tissue Microarrays

	Protein	Location of staining	Normal	Negative	Weak	Moderate	Strong
cDNA array, cluster I gene-encoded	ANG	Cytoplasm + Stroma	(+)	17	19	5	14
	BCL2	Cytoplasm	(+)	23	10	17	5
	ESR1/ER	Nucleus	(+)	22	14	5	14
	GATA3	Nucleus	(+)	26	12	7	10
	MUC1	Cytoplasm	(+)	3	19	8	25
	THBS1	Cytoplasm + Stroma	(+)	16	30	9	0
cDNA array, cluster II gene-encoded	MLANA	Cytoplasm	(+)	22	20	6	7
	PRLR	Membrane	(+)	21	12	7	15
Others	CDH1	Membrane	(+)	6	10	14	25
	ERBB2 (CB 11)	Membrane	(−)	34	14	4	3
	ERBB2 (AO485)	Membrane	(−)	30	11	9	5
	ERBB2 (3B5)	Membrane	(−)	37	8	2	8
	FGFR1	Membrane + Cytoplasm	(+)	20	20	12	3
	Ki67	Nucleus	(+)	4	27	13	11
	P53	Nucleus	(−)	33	12	0	10
	TACC1	Cytoplasm	(+)	21	22	9	3
	TACC2	Cytoplasm	(+)	5	18	13	19

(+) and (−) mean expressed or not in normal breast tissue, respectively.

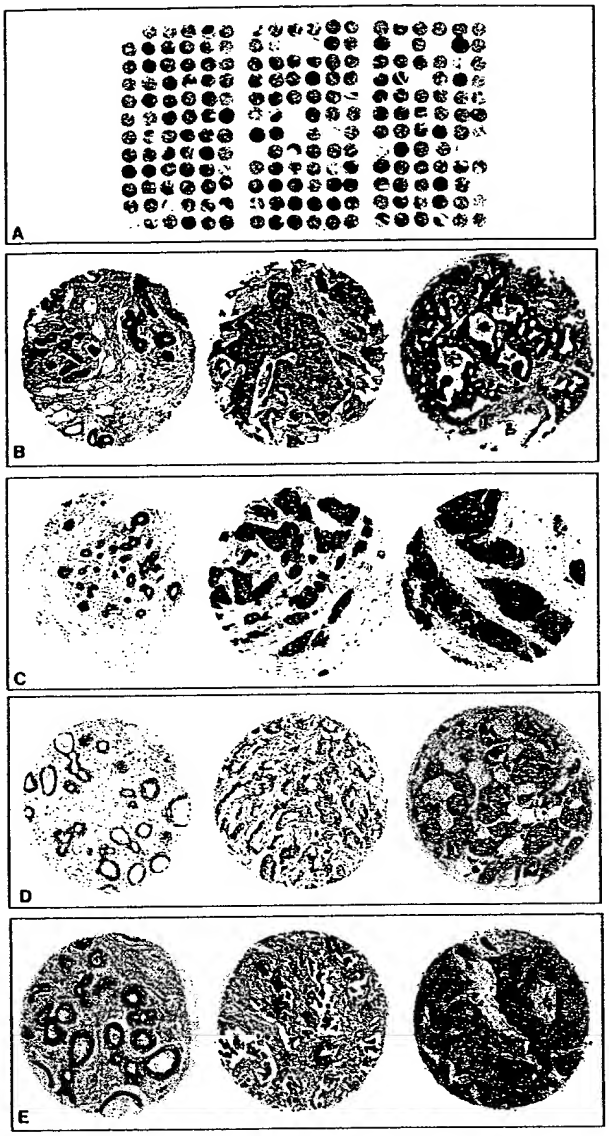


Figure 2. Expression of proteins studied by IHC on TMAs. **A:** H&E staining of a paraffin block section (25 × 30 mm) from the TMA containing 216 arrayed tumor (3 × 55) and control samples. **B:** Anti-angiogenin staining. **C:** Anti-FGFR1 staining. **D:** Anti-GATA3 staining. **E:** Anti-PRLR staining. From **B** to **E**, the first section is from normal breast tissue, the second and third from tumor tissue (the second illustrates a moderate staining whereas the third illustrates a strong staining). Original magnifications, ×50.

Comparison of the Results Obtained by cDNA Arrays and TMAs

Expression levels obtained by IHC on TMA and by cDNA array hybridizations were compared for the 15 molecules. Data from TMA analyses are discontinuous, whereas those obtained by cDNA array analyses are continuous. To facilitate comparisons, we transformed the cDNA array values into discontinuous data. Tumors were then grouped into two or three classes for each method (Table 4). Homogeneous classes were defined for TMA, by grouping tumors with an equivalent staining level (see Table 3). For cDNA arrays, classes were visually defined on examination of the distribution graphs (Figure 3).

Each tumor sample was then placed into one of the three TMA classes and attributed 1, 2, or 3, and into one of the three cDNA array classes and attributed 1, 2, or 3. Table 4 shows the number of samples in each class. Concordance between the two scores was evaluated by a contingency table analysis. A strong concordance was seen for 5 of the 15 comparisons with similar expression levels measured by the two methods: ER, ERBB2, and GATA3 ($P < 0.001$), BCL2 ($P < 0.02$), and TACC1 ($P < 0.05$). No concordance was seen for ANG, CDH1, FGFR1, Ki67, MLANA, MUC1, P53, PRLR, TACC2, and THBS1. Figure 4 shows example of comparative graphs.

Groups of Co-Regulated Molecules

Using cDNA arrays and hierarchical clustering, we had evidenced a co-expression of *ESR1* (encoding ER-α), *BCL2*, and *GATA3* at the mRNA level in breast tumors,[1,10] with a statistically significant correlation between *ESR1* and *GATA3* ($r = 0.73$, $R^2 = 0.53$, $P < 0.0001$). As shown in Figure 5A, the correlation between the three molecules was statistically confirmed at the protein level as measured by IHC on TMA. FGFR1, TACC1, and TACC2 protein levels also varied together but the correlation was weaker (Figure 5B). For each pairwise comparison, with the same number of samples ($n = 55$), we calculated a coefficient of correlation and a P value: BCL2/ER, $r = 0.79$, $R^2 = 0.62$, $P < 0.0001$; GATA3/ER, $r = 0.74$, $R^2 = 0.54$, $P < 0.0001$; TACC1/FGFR1, $r = 0.67$, $R^2 = 0.45$, $P < 0.001$; and TACC2/FGFR1, $r = 0.57$, $R^2 = 0.32$, $P < 0.001$.

Impact on Survival of RNA and Protein Expression Levels

To further estimate the clinical interest of the cDNA array and TMA combined approach, we examined and compared the prognostic information provided by mRNA and protein expression levels for each of the 15 molecules independently. Only 2 of the 15 tested markers showed individual prognostic value. High *THBS1* mRNA levels were associated with a better survival whereas no such correlation was found with protein levels. The opposite was true for MUC1: low levels of MUC1 protein were associated with a better survival, whereas mRNA levels did not correlate with survival (Figure 6). Thus, depending on the marker, clinically relevant information was differently provided by cDNA or TMA technique, suggesting that both analyses are worth performing simultaneously on the same cases.

These results were obtained on a limited number of cases representing a selected population of poor prognosis localized tumors. We sought to confirm the observation on MUC1 on a larger series of cases (Figure 7A). We studied 592 samples (including the 55) arrayed in a second TMA with anti-MUC1 antibody. MUC1 staining in normal cells is either absent or detected in the apical membrane; tumor cells express MUC1 in two abnormal localizations (cytoplasm or circumferential membrane) and a strong cytoplasmic staining is associated with a

Table 4. Comparison of Expression Levels Measured Using Analyses of Tissue Microarrays and cDNA Arrays

Gene	Tissue microarray classes			cDNA array classes			Concordance
	1	2	3	1	2	3	*P* values
ESR1/ER	22 (N)	19 (W +M)	14 (S)	15	22	18	<0.001
BCL2	23 (N)	10 (W)	22 (M +S)	18	37	0	<0.02
P53	33 (N)	22 (W + M+S)	/	46	9	0	NS
GATA3	26 (N)	12 (W)	17 (M +S)	18	25	12	<0.001
PRLR	21 (N)	19 (W +M)	15 (S)	36	12	7	NS
ERBB2 (3B5)	37 (N)	10 (W +M)	8 (S)	46	4	5	<0.001
CDH1	16 (N +W)	14 (M)	25 (S)	42	13	0	NS
TACC2	23 (N +W)	13 (M)	19 (S)	19	18	18	NS
TACC1	21 (N)	22 (W)	12 (M +S)	19	18	18	<0.05
MLANA	22 (N)	20 (W)	13 (M +S)	34	21	0	NS
FGFR1	20 (N)	20 (W)	15 (M +S)	45	10	0	NS
ANG	17 (N)	19 (W)	19 (M +S)	17	30	8	NS
THBS1	16 (N)	30 (W)	9 (M +S)	46	6	8	NS
Ki67	31 (N +W)	24 (M + S)	/	11	33	11	NS
MUC1	22 (N +W)	33 (M + S)	/	39	8	8	NS

N, Negative; W, weak; M, moderate; S, strong; NS, not significant.
Numbers for tissue microarrays are taken from Table 3 and numbers for cDNA arrays are obtained using the method shown in Figure 3.

poor prognosis.[22] For the 55 tumors of the first TMA, the prognostic value of the quantitative quick score was related to a high frequency of abnormal cytoplasmic and circumferential MUC1 localizations (83%) as compared to apical localization and absence (17%). Of the 592 cases of the second TMA, 551 were available for analysis after MUC1 staining: 249 cases (45%) showed apical or no staining, 302 (55%) displayed cytoplasmic or circumferential membrane staining (Figure 7B). In this larger series the quantitative quick score did not have a significant prognostic value. This was because of the fact that the topographical aspect was significantly different from that of the short series with only 55% *versus* 83% of cytoplasmic and circumferential localizations. When considered, qualitative assessment of the staining provided prognostic information; the apical localization and the absence of MUC1 strongly correlated with a better evolution (*P* = 0.0154) (Figure 7C).

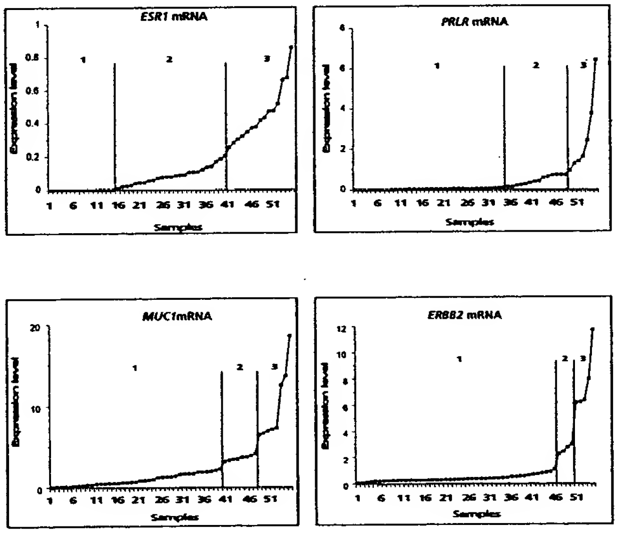


Figure 3. Transformation of continuous cDNA array data into discontinuous data. mRNA expression levels measured by cDNA array are plotted for each sample in an increasing order. For each gene, classes are determined on visual inspection and are separated by **vertical bars** on the graphs. Results for ER-α (*ESR1*), prolactin receptor (*PRLR*), mucin 1 (*MUC1*), and *ERBB2* are shown.

Discussion

The recent availability of new high-throughput molecular analyses offers the opportunity to tackle the complexity and the combinatorial nature of breast cancer at the molecular level. Expected applications are a better un-

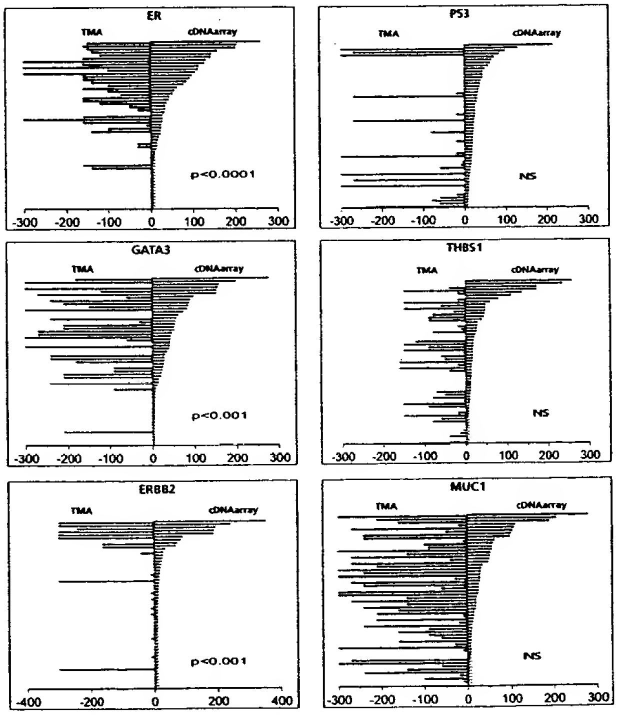


Figure 4. Comparison of data obtained by cDNA array and IHC on TMA. Results for each tumor (from top to bottom) are represented as **opposite bars**, with the value of IHC (quick score) on the **left**, and the value of the cDNA array analyses (artificially ×30 for visual effect) on the **right**. Values for ER, GATA3, and ERBB2 show good correlation between the two methods, whereas values for P53, THBS1, and MUC1 do not show such correlation.

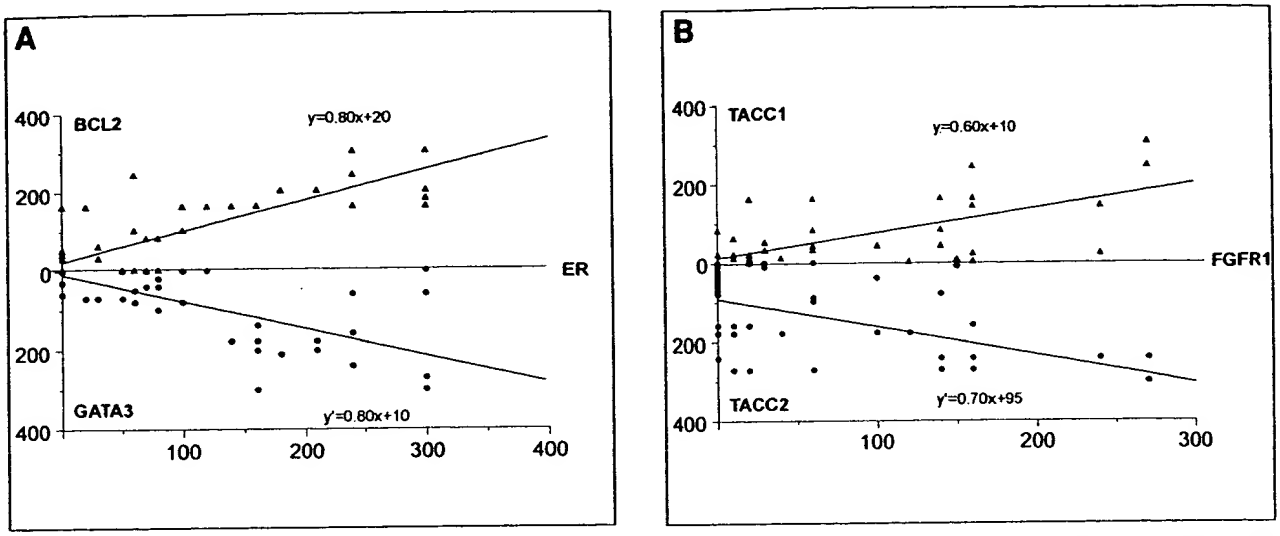


Figure 5. Similar variations in expression levels of two groups of proteins. **A:** The expression levels of ER, BCL2, and GATA3 as measured by IHC on TMAs correlated, as determined by simple linear regression analysis. **B:** Similarly, the expression levels of FGFR1, TACC1, and TACC2 correlated.

derstanding of the disease and the identification of new diagnostic and prognostic markers and therapeutic targets, both needed to improve the management of patients. At the same time it introduces a new challenge for pathologists who, in charge of the first assessment of the tumors, need to know how to optimally use these new methods. The present study directly followed a cDNA array-based analysis of a breast tumor series. The tumor samples were obtained from 55 women with poor prognosis breast cancer treated with adjuvant chemotherapy. Currently such patients have a long-term survival of ~70% and there is a crucial need to identify parameters that might accurately predict the clinical outcome in individual patients. Our study was designed to evaluate the interest and limitations of IHC on TMA as a natural extension of the cDNA array approach in a hospital setting.

We first confronted cDNA array and TMA analyses to other methods, ie, RQ-PCR and conventional IHC, respectively. The good concordance between mRNA expression levels observed by cDNA arrays and RQ-PCR further confirmed the validity of our cDNA array measurements. TMAs allow to screen large series of tumor samples using several archival materials, but their representation of the entire tumor has been questioned. Our degrees of concordance between stainings on full sections and on TMA were in the same range as published studies. Several authors have reported that TMA constructed with three cores per sample (as in our study) are representative of whole tumor specimens.[23–30]

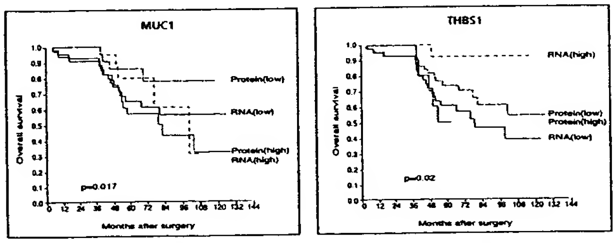


Figure 6. Kaplan-Meier plots of patient overall survival. **Left:** Survival according to MUC1 mRNA and protein expression levels. **Right:** Survival according to THBS1 mRNA and protein expression levels (labeled high and low). High and low protein levels correspond to strong plus moderate *versus* weak plus negative (see Table 3), respectively, and high and low mRNA levels correspond to classes 2 and 3 *versus* class 1 (see Figure 4), respectively.

As a large-scale validation tool of DNA or RNA data, IHC on TMAs should be interpreted with caution. Indeed, comparison of our cDNA array and TMA data, obtained on the same breast tumor samples, gave different results according to the gene product examined.

For a category of molecules we found important differences between RNA and protein expression levels. This was the case of P53. This discrepancy was rather expected because P53 protein detection is not dependent on mRNA overexpression, but is because of the increased half-life of a mutated protein. In normal cells, P53 protein half-life is short and expression levels are low and undetectable by IHC. In cancer cells, most P53 mutations lead to products that are not ubiquitinated and accumulate in the nuclei where they can then be detected. Other noteworthy cases were MUC1 and THBS1. These differences certainly stem from the fact that different levels of biological information are examined. For many genes, there is little correlation between the abundance of the mRNA transcript with steady-state levels of the encoded protein. Posttranscriptional and posttranslational mechanisms are likely to influence protein expression, thus blurring the correlation between mRNA and protein levels. Proteins encoded by very low levels of RNA, ie, below the detection level of cDNA arrays, can be detected by IHC because of increased protein stability (eg, the case of P53) or high sensitivity of the antibody, and reciprocally, elevated levels of RNA may produce only little amounts of detectable proteins. Special calibration of the antibody aimed to detect only a certain level of protein is another limitation. The chosen antibody may also detect only certain forms of a protein that do not correspond to the cDNA spotted on the DNA array, because of alternative splicings of mRNA for example. This particularly can explain the difference observed between THBS1 mRNA and protein levels, and conse-

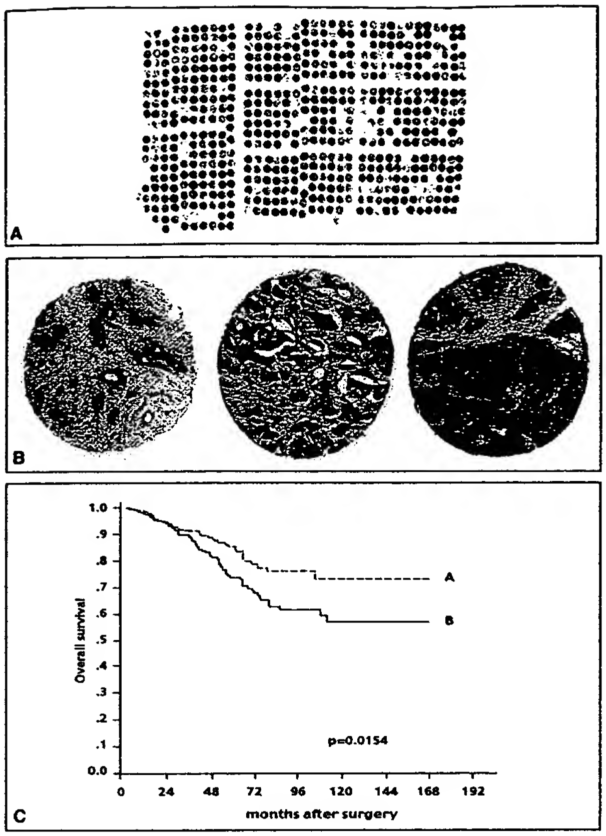


Figure 7. Expression of MUC1 protein studied by IHC on a tissue-microarray. **A:** H&E staining of a paraffin block section (25 ×. 30 mm) from the TMA containing 647 arrayed samples, including 592 tumors and 55 controls. **B:** MUC1 staining: normal breast tissue (**left**), apical (**middle**), and cytoplasmic (**right**) staining in tumors. **C:** Kaplan-Meier plot of patient overall survival: survival differs significantly according to MUC1 protein localization. **A:** Absence of staining or apical localization; **B:** cytoplasmic or circumferential membrane localization.

quently their different prognostic impact.[31] Finally, distinct areas of a heterogeneous tumor may be submitted to RNA and protein analyses.

Conversely, we observed an excellent correlation between RNA and protein levels in one-third of the tested molecules. This was the case for ERBB2, despite the fact that its corresponding antibody is calibrated to detect only overexpression. Among the other molecules with correlated mRNA and protein expression levels were ER, GATA3, and BCL2. We and others had shown that the mRNA levels of the three genes covaried in cDNA array analyses.[1,10,32] Here we were able to confirm this coexpression at the protein level. This group of co-regulated genes and proteins may be linked to the hormonal control of the mammary gland. Such identification is important for a better understanding of gene and protein networks that operate in cancer cells; it may lead to the discovery of new molecules to be targeted to block or stimulate a metabolic pathway or function; it may also provide a prognostic information clinically more relevant than that of isolated markers because it better reflects the functional status of a pathway such as the estrogen pathway of breast tumors.

Several studies have shown the interest of TMA studies in cancer research to extend cDNA array data.[33] A pioneering analysis was conducted by Moch and colleagues;[8] after the identification of vimentin as overexpressed in a renal cancer cell line using cDNA arrays, the authors extended this result to the protein level on a series of 532 tumor specimens arrayed onto a renal cancer TMA. Using TMA of bladder tumors containing 2317 specimens from 1842 patients, Richter and colleagues[9] found a positive correlation between *CCNE* gene amplification measured by fluorescence *in situ* hybridization and cyclin-E protein overexpression measured by IHC. The combination of cDNA array and TMA allowed the identification of IGFBP2 and HSP27,[34] hepsin,[2] and AMACR[30] as significantly overexpressed in prostate cancer, suggesting their putative diagnostic interest. IGFBP2 was also found as a marker of poor prognosis in a series of 418 brain tumors arrayed onto a TMA.[35] A similar study showed the overexpression of the WT protein in ovarian cancer.[36] The expression level of PKCβ was measured by IHC on a B-cell lymphoma TMA to validate cDNA array data.[3] In breast carcinomas, Hedenfalk and colleagues[37] showed that, like mRNA levels, cyclin D1 protein levels were differentially distributed among BRCA1 and BRCA2 hereditary tumors. All these studies showed a good correlation between the two techniques of investigation, but were limited to the analysis of a single highly selected marker and were not, with few exceptions, conducted on the same samples. Our present study is the first deliberate comparative analysis of cDNA and TMAs. It shows a correlation between the two techniques for one-third of the selected markers and the absence of correlation for the other two-thirds.

These discrepancies deserve two commentaries. First, given the flurry of encouraging data associated with the rapidly emerging cDNA array technology, it is paramount to determine to what extent changes in mRNA expression are accompanied or not by similar changes at the protein level. In some cases, the differences may be eliminated by a number of experimental precautions, such as selection of biopsy cores and antibodies, but in other cases, they will remain. If protein levels of a target molecule, or a group of molecules, correlate with its selection by cDNA array, IHC on TMA offers a powerful tool to quickly evaluate the clinical relevance of differentially expressed genes. But if they do not correlate, the cDNA array and TMA results must be considered independently because each can provide distinct information.

Second, even if the intrinsic prognostic power of cDNA array data and clustering analyses derives from the combined expression of several genes, and not from an individual gene, it may be interesting for routine clinical application to test each of these genes as a candidate marker and to determine how its expression may alone distinguish the tumor classes. The main interest of TMA lies in the possibility to test large series of tumor samples with individual markers. In our series of samples, we observed that mRNA, but not protein expression levels of *THBS1* had prognostic value, suggesting that they play an important role in the discriminator power of the cDNA array gene cluster. In contrast, for MUC1, as seen earlier,[38] low levels of protein were associated with a better prognosis, which was not the case for mRNA; IHC further allowed a qualitative appreciation of the protein localization, which happened to be crucial information for prognosis when an unselected population was studied.

In the period of validation studies that has now begun, for which retrospective IHC studies on archival paraffin-embedded material are required,[6] it is particularly important to bear in mind that differences between mRNA and protein expression levels are possible with respect to intensities and to prognostic relevance. These differences underline the complementarity or synergy between expression measurements from cDNA arrays and IHC on TMA, and also the need for other high-throughput technologies such as cDNA arrays containing alternatively spliced transcripts,[39] protein arrays,[40] and *in situ* hybridizations on TMAs.[41] The combination of these complementary approaches will accelerate even more the identification of new diagnostic and prognostic markers as well as new therapeutic targets and will improve the management of breast cancer patients.

Acknowledgments

We thank J. Hassoun, G. Houvenaegel, D. Maraninchi, and C. Mawas for encouragement; and I. Bièche (Unité propre de recherche et d'enseignement supérieur JE2195, Paris, France), J. M. Durey (Institut Paoli-Calmettes, Marseille, France), O. P. Kallioniemi (National Human Genome Research Institute, Bethesda, MD), S. Lakhani (The Breakthough Toby Robins Breast Cancer Research Centre, London, UK), R. Lidereau (E0017 INSERM, St-Cloud, France), and B. Puig (Institut Paoli-Calmettes, Marseille, France) for help and advice.

References

1. Bertucci F, Houlgatte R, Benziane A, Granjeaud S, Adelaide J, Tagett R, Loriod B, Jacquemier J, Viens P, Jordan B, Birnbaum D, Nguyen C: Expression profiling in primary breast carcinomas using arrays of candidate genes. Hum Mol Genet 2000, 9:2981–2991
2. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pientas KJ, Rubin MA, Chinnaiyan AM: Delineation of prognostic biomarkers in prostate cancer. Nature 2001, 412:822–825
3. Shipp MA, Ross KN, Tamayo P, Weng A, Kutok JL, Aguiar RCT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 2002, 8:68–74
4. Van't Veer LJ, Dai H, van de Vijver M, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002, 415:530–535
5. Bertucci F, Houlgatte R, Nguyen C, Viens P, Jordan B, Birnbaum D: Gene expression profiling of cancer using DNA arrays: how far from the clinic? Lancet Oncol 2001, 2:674–682
6. Lakhani S, Ashworth A: Microarray and histopathological analysis of tumours: the future and the past? Nat Rev Cancer 2001, 1:151–157
7. Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP: Tissue microarrays for high-throughput molecular profiling of tumors specimens. Nat Med 1998, 4:844–847
8. Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch MJ, Kallioniemi OP, Sauter G: High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. Am J Pathol 1999, 154:981–986
9. Richter J, Wagner U, Kononen J, Fijan A, Bruderer J, Schmid U, Ackerman D, Maurer R, Alund G, Knönagel H, Rist M, Wilber K, Anabitarte M, Hering F, Hardmeier T, Schönenberger A, Flury R, Jäger P, Fehr JL, Schrami P, Moch H, Mihatsch MJ, Gasser T, Kallioniemi OP, Sauter G: High-throughput tissue microarray analysis of cyclin E gene amplification and overexpression in urinary bladder cancer. Am J Pathol 2000, 157:787–794
10. Bertucci F, Nasser V, Granjeaud S, Eisinger F, Tagett R, Adélaïde J, Loriod B, Benziane A, Giaconia A, Devilard E, Jacquemier J, Viens P, Nguyen C, Birnbaum D, Houlgatte R: Gene expression profiles of poor prognosis primary breast cancer correlate with survival. Hum Mol Genet 2002, 11:863–872
11. Möbus VJ, Moll R, Gerharz CD, Kieback DG, Merk O, Runnebaum IB, Linner S, Dreher L, Grill HJ, Kreienberg R: Differential characteristics of two new tumorigenic cell lines of human breast carcinoma origin. Int J Cancer 1998, 77:415–423
12. Theillet C, Adélaïde J, Louason G, Bonnet-Dorion F, Jacquemier J, Adnane J, Longy M, Katsaros D, Sismondi P, Gaudray P, Birnbaum D: FGFR1 and PLAT genes and DNA amplification at 8p12 in breast and ovarian cancers. Genes Chromosom Cancer 1993, 7:219–226
13. Bertucci F, Van Hulst S, Bernard K, Loriod B, Granjeaud S, Tagett R, Starkey M, Nguyen C, Jordan B, Birnbaum D: Expression scanning of an array of growth control genes in human tumor cell lines. Oncogene 1999, 18:3905–3912
14. Bertucci F, Bernard K, Loriod B, Chang YC, Granjeaud S, Birnbaum D, Nguyen C, Peck K, Jordan BR: Sensitivity issues in DNA array-based expression measurements: advantages of Nylon microarrays for small samples. Hum Mol Genet 1999, 8:1715–1722
15. Bièche I, Franc B, Vidaud D, Vidaud M, Lidereau R: Analyses of MYC, ERBB2 and CCND1 genes in benign and malignant thyroid follicular cell tumors by real-time polymerase chain reaction. Thyroid 2001, 11:147–152
16. Mitas M, Mikhitarian K, Walters C, Baron PL, Elliott BM, Brothers TE, Robison JG, Metcalf JS, Palesch YY, Zhang Z, Gillanders WE, Cole DJ: Quantitative real-time RT-PCR detection of breast cancer micrometastasis using a multigene marker panel. Int J Cancer 2001, 93:162–171
17. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lonning PE, Borresen-Dale AL: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001, 98:10869–10874
18. Ahr A, Karn T, Solbach C, Seiter T, Strebhardt K, Holtrich U, Kaufmann M: Identification of high risk breast-cancer patients by gene expression profiling. Lancet 2002, 359:131–132
19. Bertucci F, Eisinger F, Houlgatte R, Viens P, Birnbaum D: Gene expression profiling of breast cancer and prognosis. Lancet 2002, 360:173–174
20. Conte N, Charafe-Jauffret E, Delaval B, Adélaïde J, Ginestier C, Geneix J, Isnardon D, Jacquemier J, Birnbaum D: Carcinogenesis and translational controls: TACC1 is down-regulated in human cancers and associates with mRNA regulators. Oncogene 2002, 21:5619–5630
21. Ugolini F, Adélaïde J, Charafe-Jauffret E, Nguyen C, Jacquemier J, Jordan B, Birnbaum D, Pébusque MJ: Differential expression assay of chromosome arm 8p genes identifies Frizzled-related (FRP1/FRZB) and fibroblast growth factor receptor 1 (FGFR1) as candidate breast cancer genes. Oncogene 1999, 18:1903–1910
22. Rahn JJ, Dabbagh L, Pasdar M, Hugh JC: The importance of MUC1 cellular localization in patients with breast carcinoma. Cancer 2001, 91:1973–1982
23. Hoos A, Cordon-Cardo C: Tissue microarray profiling of cancer specimens and cell lines: opportunities and limitations. Lab Invest 2001, 81:1331–1338
24. Nocito A, Kononen J, Kallioniemi OP, Sauter G: Tissue microarrays (TMAs) for high-throughput molecular pathology research. Int J Cancer 2001, 94:1–5
25. Rimm DL, Camp RL, Charette LA, Olsen DA, Provost E: Amplification

of tissue by construction of tissue microarrays. Exp Mol Pathol 2001, 70:255–264

26. Camp RL, Charette LA, Rimm DL: Validation of tissue microarray technology in breast carcinoma. Lab Invest 2000, 80:1943–1949
27. Hoos A, Urist MJ, Stojadinovic A, Mastorides S, Dudas ME, Leung DHY, Kuo D, Brennan MF, Lewis JL, Cordon-Cardo C: Validation of the tissue microarray for immunohistochemical profiling of cancer specimens using the example of human fibroblastic tumors. Am J Pathol 2001, 158:1245–1251
28. Torhorst J, Bucher C, Kononen J, Haas P, Zuber M, Kochli OR, Mross F, Dieterich H, Moch H, Mihatsch M, Kallioniemi OP, Sauter G: Tissue microarray for rapid linking of molecular changes to clinical endpoints. Am J Pathol 2001, 159:2249–2256
29. Rubin MA, Dunn R, Strawderman M, Pienta KJ: Tissue microarray sampling strategy for prostate cancer biomarker analysis. Am J Surg Pathol 2002, 26:312–319
30. Rubin MA, Zhou M, Dhanasekaran SM, Varambally S, Barrette TR, Sanda MG, Pienta KJ, Ghosh D, Chinnaiyan AM: Alpha-methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. JAMA 2002, 287:1662–1670
31. Matthias LJ, Gotis-Graham I, Underwood PA, McNeil HP, Hogg PJ: Identification of monoclonal antibodies that recognize different disulfide bonded forms of thrombospondin 1. Biochem Biophys Acta 1996, 1296:138–144
32. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA 1999, 96:9212–9217
33. Mousses S, Kallioniemi A, Kauraniemi P, Elkahloun A, Kallioniemi OP: Clinical and functional target validation using tissue and cell microarrays. Curr Opin Chem Biol 2001, 6:97–101
34. Bubendorf L, Kolmer M, Kononen J, Koivisto P, Mousses S, Chen Y, Mahlamaki E, Schraml P, Moch H, Willi N, Elkahloun AG, Pretlow TG, Gasser TC, Mihatsch MJ, Sauter G, Kallioniemi OP: Hormone therapy failure in human prostate cancer: analysis by complementary DNA and tissue microarrays. J Natl Cancer Inst 1999, 91:1758–1764
35. Sallinen SL, Sallinen PK, Haapasalo HK, Helin HJ, Helen PT, Schraml P, Kallioniemi OP, Kononen J: identification of differentially expressed genes in human gliomas by DNA microarray and tissue chip techniques. Cancer Res 2000, 60:6617–6622
36. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF, Hampton GM: Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 2001, 61:7388–7393
37. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J: Gene-expression profiles in hereditary breast cancer. N Engl J Med 2001, 344:539–548
38. McGuckin MA, Walsh MD, Hohn BG, Ward BG, Wright RG: Prognostic significance of MUC1 epithelial mucin expression in breast cancer. Hum Pathol 1995, 26:432–439
39. Yeakley JM, Fan JB, Doucet D, Luo L, Wickham E, Ye Z, Chee MS, Fu XD: Profiling alternative splicing on fiber-optic arrays. Nat Biotechnol 2002, 20:353–358
40. Emili AQ, Cagney G: Large-scale functional analysis using peptide or protein arrays. Nat Biotechnol 2000, 18:393–397
41. Fejzo MS, Slamon DJ: Frozen tumor tissue microarray technology for analysis of tumor RNA, DNA, and proteins. Am J Pathol 2001, 159: 1645–1650

NCBI PubMed National Library of Medicine NLM

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Bc

Search PubMed for Go Clear

☑ Limits Preview/Index History Clipboard Details

About Entrez

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby

Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

Privacy Policy

Display Abstract Show: 20 Sort Send to Text

□ 1: Int J Mol Med. 1998 May;1(5):855-61. Related Articles, Lin

p185 overexpression in 220 samples of breast cancer undergoing primary surgery: comparison with c-erbB-2 gene amplification.

Dalifard I, Daver A, Goussard J, Lorimier G, Gosse-Brun S, Lortholary A, Larra F.

Laboratoire de Radioanalyse, Centre Paul Papin, 49033 Angers Cedex 01, France.

In breast cancer, DNA amplification of the oncogene c-erbB-2, encoding for the p185 protein, is associated with a poor prognosis. A retrospective study o a population of 220 cases of primary breast cancer permitted a quantitative measure of p185 oncoprotein overexpression by an immunoenzymetric assay and the determination of c-erbB-2 amplification by the Southern blot method A correlation existed between the two measurements (r=0.85) using the double cut-off: DNA 2 copies and p185 400 U/mg protein, and only 2.7% of the cases were discordant. 13.2% of the tumors showed p185 overexpression The percentage of tumors overexpressing p185 was significantly different between the groups with amplified and non-amplified c-erbB-2. We observe a significant correlation between p185 levels and tumor grade (p=0.03), and an inverse correlation with hormonal receptors (p=0.0001). The p185 assay could be an additional prognostic factor to better define patient subgroups with node negative, grade II, and positive or negative hormonal receptors.

PMID: 9852307 [PubMed - indexed for MEDLINE]

Display Abstract Show: 20 Sort Send to Text

Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services
Freedom of Information Act | Disclaimer

Jun 5 2003 10:08

Eur. J. Biochem. 189, 475 – 486 (1990)
© FEBS 1990

A transcribed gene, containing a variable number of tandem repeats, codes for a human epithelial tumor antigen

cDNA cloning, expression of the transfected gene and over-expression in breast cancer tissue

Mara HAREUVENI[1, 2], Ilan TSARFATY[1], Joseph ZARETSKY[1], Phillip KOTKES[1], Judith HOREV[1], Sheila ZRIHAN[1], Mordechai WEISS[1], Stephen GREEN[2], Richard LATHE[2], Iafa KEYDAR[1] and Daniel H. WRESCHNER[1]

[1] Department of Microbiology, Faculty of Life Sciences, Tel Aviv University, Israel
[2] Laboratoire de Genetique Moleculaire des Eucaryotes du Center National Récherche Scientifique, Institut de Chimie Biologique, Faculte de Medecine, Strasbourg, France

(Received July 28/December 13, 1989) – EJB 89 0945

A monoclonal antibody, H23, that specifically recognizes a breast-tumor-associated antigen, was used to isolate a cDNA insert that codes for the antigenic epitope. Nucleotide sequencing of this cDNA, as well as a longer 850-bp cDNA insert, shows that they are composed of 60-bp (G + C)-rich tandem repeating units. The coding strand was determined and codes for a proline-rich 20-amino-acid repeat motif. A comparison of the highly conserved repeat unit with the deduced flanking amino acid sequences demonstrates conservation of specific subregions of the repeat consensus within the flanking amino acids. Hybridization of the 60-bp cDNA probe with RNAs extracted from a variety of primary and metastatic human tumors yields relatively high levels of hybrid with the breast carcinomas, as compared to lower hybrid levels with RNAs from other epithelial tumors. RNA extracted from breast tissue adjacent to the tumor or from benign breast tumors, demonstrates low or undetectable levels of hybridization. Probing Southern blots with the 60-bp repeat shows that the tumor antigen is highly polymorphic and contains a variable number of tandem repeats (VNTRs). The VNTR nature of the gene was confirmed by probing Southern blots with unique genomic sequences that are physically linked to an isolated gene fragment that also contains the tandem repeat array. Mouse cells transfected with this gene fragment produce tumor antigen that is readily detected by H23 monoclonal antibodies. The allelic forms seen in 10 different primary human tumors demonstrate 100% concordance with the various mRNA species expressed. These studies are extended to the protein forms detected by immunoblot analyses that show both a correlation of the expressed tumor antigen species with the allelic forms as well as significantly increased expression in breast cancer tissue. The above studies unequivocally establish the over-expression of a VNTR gene coding for an epithelial tumor antigen in human breast cancer tissue.

The isolation and characterization of proteins that are aberrantly expressed in human tumor tissues may elucidate cellular mechanisms leading to malignancy and also be of significant clinical importance. To identify breast-tumor-associated markers, we have established a human breast cancer cell line, T47D, that has been extensively studied and retains many characteristics of primary human breast tumors [1]. Monoclonal antibodies (mAb) were prepared against particulate antigens released by these cells and screened for their specificity by the immunohistochemical staining of breast tissue sections. One mAb, designated H23, stained cytoplasmically 91% of all malignant breast tumors analyzed, whereas little or no cytoplasmic staining was observed in normal and benign breast tissues [2].

Other groups have also described mAbs reactive with high-molecular-mass glycoproteins that are aberrantly expressed in epithelial tumors and especially in breast cancer [3 – 12]. Several of these mAbs, namely DF3, HMFG-1, HMFG-2 and SM-3, were used to isolate cDNA clones that express the immunereactive epitope [13 – 15]. The cDNAs isolated all contain tandem 60-bp repeat units [13 – 15] that code for a 20-amino-acid repeat motif rich in proline, serine, threonine and alanine [14]. Southern blots probed with the 60-bp repeat cDNA insert show that the gene is highly polymorphic and correlates with the polymorphism observed in the protein products [13, 16]. These results suggest the codominant autosomal expression of a gene that contains a variable number of tandem repeats (VNTR).

Besides the 60-bp repeat unit, little has been known regarding the unique non-repeat sequences of the tumor antigen cDNA and its gene. Indeed, the aforementioned studies on the molecular structure of the epithelial antigen, including all Southern and Northern blot analyses, have been performed solely with the 60-bp cDNA repeat unit [13 – 16]. To study the breast-tumor-associated antigen recognized by H23 mAbs (H23 Ag) on a molecular level, a gt11 cDNA expression library prepared from T47D mRNA was screened with these

Correspondence to D. H. Wreschner, Department of Microbiology, Tel Aviv University, Ramat Aviv, I-69978, Israel

Abbreviations. mAb, monoclonal antibody; H23 Ag, epithelial tumor antigen recognized by H23 monoclonal antibodies; ORF, open reading frame; VNTR, variable number of tandem repeats.

Note. The novel nucleotide sequence data published here and in the preceding paper in this journal have been deposited with the EMBL sequence data bank and are available under the accession number X52228 and X52229. The novel amino acid sequence data have been deposited with the EMBL sequence data bank.

mAbs. A cDNA insert that codes for the immunereactive epitope was isolated and sequenced. This insert, as well as a longer 850-bp cDNA insert, are composed of 60-bp tandem repeats, similar to those previously reported [13 – 16]. We have recently increased our knowledge beyond the confines of the 60-bp repeat units by isolating almost full-length cDNAs that contain unique non-repeat sequences located 5′ and 3′ to the tandem repeat array and code for the complete epithelial tumor antigen [16a].

We have extended these studies and report here the determination of the VNTR nature of the gene by analyzing Northern and Southern blots with probes consisting not only of the 60-bp cDNA repeat [13 – 16] but also with probes derived from unique non-repeat genomic sequences. These investigations were performed on nucleic acids isolated from primary human tissues and are therefore relevant to the *in-vivo* situation. In addition, we demonstrate the over-expression of mRNA coding for the tumor antigen and of the antigen itself in primary breast cancer tissues. The coding strand of the 60-bp repeat unit has been determined and a comparative analysis of the tumor antigen unique non-repeat amino acid sequences with the 20-amino-acid repeat motif is presented. Finally, by transfecting cells with the isolated gene coding for the tumor antigen, stable mouse cell transfectants have been established that express the human-breast-tumor-associated antigen.

The findings reported here unequivocally establish the over-expression in human breast cancer tissue of a VNTR gene that codes for an epithelial tumor antigen.

MATERIALS AND METHODS

Plating of the recombinant cDNA phage and library screening with H23 mAb

The randomly primed λgt11 cDNA expression library [17] was prepared from poly(A)-rich RNA of T47D cells, a human breast carcinoma cell line [2]. Approximately 10^6 phages were plated on *Escherichia coli* strain Y1090 and the resulting plaques were screened for expression of crossreacting galactosidase fusion protein, with 25 μg/ml H23 IgG as described elsewhere [18]. For the final detection of positive plaques, ^{125}I-protein A was used at a final concentration of 4×10^5 cpm/ml. Positive plaques were picked and rescreened repeatedly until all plaques were immunopositive. Most of the λ cDNA clones contained an insert of similar size and the clone with the longest insert, designated 3b, was thus obtained and used for Northern hybridization assays.

DNA hybridization of cDNA library

The cDNA library replica-plated on nylon membranes (Amersham, England) was probed with cDNA inserts labelled by nick translation [19] to a specific activity $2 - 5 \times 10^8$ cpm/μg and a final concentration of $1 - 2 \times 10^6$ Cerenkov cpm/ml. The replica blots were prehybridized and probed at 42°C for 15 h in 50% formamide, 5 × NaCl/Cit (1 × NaCl/Cit is 150 mM NaCl, 15 mM sodium citrate, pH 7.0), 0.1% polyvinylpyrrolidone, 0.1% Ficoll, 0.2% SDS and 100 μg/ml denatured salmon sperm DNA.

Following hybridization, the blots were washed at 65°C for 2 – 4 h with several changes of 2 × NaCl/Cit, 0.2% SDS following by stringent washing at 65°C (2 × 30 min) with 0.2 × NaCl/Cit, 0.5% SDS. The washed blots were exposed to Agfa Gevaert Curix X-ray films at – 70°C using curix-special intensifying screens.

Southern blot DNA analysis

High-molecular-mass DNA was isolated from powdered surgically removed frozen (– 70°C) tissues by incubating overnight at 50°C in 200 μg/ml proteinase K, 100 mM NaCl, 10 mM Tris/HCl pH 7.5, 1 mM EDTA followed by phenol/chloroform and one chloroform extraction. The DNA was spooled onto glass rods after the addition of 0.2 M NaCl (final concentration) and 1 vol. absolute ethanol at – 20°C. The spooled DNA was rinsed with 70% ethanol, briefly dried, resuspended in double distilled water and kept at – 20°C. The DNA (100 μg/ml final concn) was incubated with the appropriate restriction enzymes (approximately 5 units enzyme/μg DNA) overnight at 37°C followed by ethanol precipitation at – 20°C. When double digestions were performed, DNA was incubated with one enzyme followed by ethanol precipitation, resuspension and then digestion with the second enzyme. 10 – 20 μg restricted DNA was electrophoresed on 0.8% agarose gels in recirculating Tris/acetate/EDTA buffer, followed by staining with ethidium bromide and washing in 1.5 M NaCl, 0.5 M NaOH for 30 min. Southern transfer to nylon membranes (Amersham, England) was performed in 1.5 M NaCl, 0.25 M NaOH. The blot was irradiated with ultraviolet light for 3 min, followed by baking at 80°C for 2 h. Prehybridization, hybridization and washing were as described above.

RNA analysis

RNA was extracted from surgically removed frozen tissues using the guanidinium thiocyanate/cesium chloride method [20]. Poly(A)-rich RNA was purified by oligo(dT)-cellulose chromatography [21]. For dot-blot analysis, 15 μg of each sample of total RNA was applied with gentle vacuum in 200 μl of 2 × NaCl/Cit to a Gelman nylon membrane using the BRL dot-blot apparatus. The RNA samples were covalently attached to the nylon membrane by ultraviolet irradiation followed by baking at 80°C under vacuum.

For Northern analysis 40 μg of each total RNA sample or 4 μg of poly(A)-rich selected RNA were subjected to electrophoresis on a 1.4% agarose gel under glyoxal/dimethylsulfoxide-denaturing conditions using Tris/acetate/EDTA as the running buffer. Subsequent to 50 mM NaOH treatment and washings in 2 × NaCl/Cit, the gels were stained by ethidium bromide and Northern blotted to Gelman nylon membranes [21].

Northern and RNA dot-blot hybridizations

The blots obtained as previously described, were prehybridized and probed at 42°C for 16 h in 50% formamide, 5 × NaCl/Cit, 0.1% polyvinylpyrrolidone, 0.1% Ficoll, 0.2% SDS and 100 μg/ml denatured salmon sperm DNA with cDNA inserts labelled by nick translation [19] to a specific activity of $2 - 5 \times 10^8$ cpm/μg. A final concentration of $1 - 2 \times 10^6$ Cerenkov cpm/ml was used. Following hybridization, the blots were washed at 65°C for 2 – 4 h with several changes of 2 × NaCl/Cit, 0.2% SDS followed by stringent washing at 65°C (2 × 30 min) with 0.2 × NaCl/Cit, 0.5% SDS. Quantification of the hybridization intensity was performed with the LKB 2222-020 Ultrascan XL II laser densitometer. Bound probes were removed by washing blots in hybridization buffer

70°C for 60 min and the membranes were then rehybridized with a different probe under similar conditions.

Construction of eukaryotic expression vector coding for H23 Ag

The *Xmn*I – *Eco*RI genomic fragment (see Fig. 6) was inserted into the eukaryotic expression vector pCL642 (this vector will be described in detail in a separate publication). Briefly, pCL642 is composed of the promoter region (1.4 kb) isolated from the mouse housekeeping gene coding for 3-hydroxy-3-methylglutaryl-coenzyme-A reductase. The promoter is followed by the untranslated first exon and intron (0.7 kb and 3.5 kb) derived also from this reductase gene. The *Xmn*I site of the *Xmn*I – *Eco*RI genomic fragment coding for the tumor antigen was blunt-end-ligated to an *Eco*RV site located in a polylinker immediately downstream to the reductase intron. The *Eco*RI site was ligated to the *Kpn*I site of the polylinker via an *Eco*RI – *Kpn*I adaptor. A 123-bp fragment containing the SV40 poly(A) signal sequence is situated immediately 3′ to the polylinker. The construct pCL642/*Xmn*I – *Eco*RI (10 μg) was contransfected with 1 μg pAG60(G418R) plasmid [22] into either MM5tC3H cells (from American Tissue Culture Collection) or FR3T3 ras-1 cells [23] using a modification [24] of the calcium phosphate precipitation method [25]. Cells were selected for G418 resistance (Geniticin 500 μg/ml) and loci were picked and subcultured.

For the detection of tumor antigen, the transfected cells were grown on coverslips and immunohistochemically stained with H23 mAbs. Control cells were either transfected with the pAG60 plasmid alone or with an irrelevant gene.

Nucleotide sequencing

Sequencing was accomplished using the dideoxynucleotide chain-termination method [26]. Restriction fragments of the cDNA inserts were subcloned into M13 and both strands were sequenced. The ssDNA was primed with either the M13 universal primer or synthetic oligonucleotides prepared according to known sequences. The analysis of the sequence was performed using the Beckman MicroGenie program.

Radioactive labelling of DNA probes

Double-stranded DNA probes were radioactively labelled with [α-^{32}P]dCTP either by nick translation or random oligonucleotide multipriming using commercially available kits (BRL, USA, and Amersham, England, respectively). All DNA probes used here were purified inserts that were isolated by agarose gel electrophoresis. Single-stranded oligonucleotides were 5′-end labelled by incubating with [γ-^{32}P]ATP and polynucleotide kinase. All labelled probes were purified from non-incorporated nucleotide by passage through Sephadex G-100 columns.

Oligonucleotide synthesis

Oligonucleotides were prepared at the Macromolecule Synthesis Service Unit (Department of Organic Chemistry, Weizman Institute of Science) by Dr Ora Goldberg using an Automated Applied Biosystems synthesizer. Following synthesis, the oligonucleotides were electrophoretically purified on acrylamide/urea gels.

Immunoblotting

Protein samples denatured by boiling in SDS buffer containing mercaptoethanol were analyzed on 3 – 15% linear gradient SDS/acrylamide gels as previously described [27]. Following electrophoresis, the gel was equilibrated in transfer buffer (Tris/glycine) and electrotransferred for 3 h at 1 A to nitrocellulose filters. The filters were blocked in NaCl/P$_i$ (150 mM NaCl, 15 mM sodium phosphate, pH 7.0) containing 5% skimmed milk (Blotto) followed by incubation with antibody in Blotto. The filters were washed in NaCl/P$_i$ and then reacted with ^{125}I-protein A (Amersham, England) dissolved in Blotto.

Monoclonal antibodies

Monoclonal antibodies (mAbs) were prepared against particulate antigens released into the medium by T47D breast carcinoma cells, using established procedures. The monoclonal antibodies obtained were screened against paraffin-embedded sections of benign and malignant breast tissue with the immunoperoxidase-staining technique and one of the mAbs designated H23 [2] was selected and used in this study.

RESULTS

Sequence of cDNA coding for epitope recognized by H23 monoclonal antibodies

The gt11 cDNA expression libraries prepared with mRNA isolated from either T47D or MCF7 [17], both human breast carcinoma cell lines, were probed with the monoclonal antibody H23. Libraries obtained by priming poly(A)-rich RNA with oligo(dT), as well as with random nucleotide oligomers, were investigated. Recombinant clones immunoreactive with H23, were obtained at a frequency of approximately 1 in 2000 in the amplified libraries and the cDNA inserts of all clones analyzed revealed a size of approximately 220 – 240 bp. Both random oligomer-primed as well as oligo(dT)-primed libraries from the MCF7 and T47D cell lines revealed similar-sized inserts.

Nucleotide sequencing of one such representative cDNA insert, termed 3b, indicates that it is (G + C)-rich with strand preference for the G or C nucleotides. Inspection of the 225-bp sequence shows that it is composed of a 60-nucleotide tandem repeat unit which is remarkably conserved with only very few substitutions occurring between the different units (Fig. 1 A).

Longer cDNA inserts contain the same 60-bp tandem repeat unit

In order to obtain longer cDNA clones, the 3b cDNA insert was used to reprobe the library by DNA/DNA hybridization. Several recombinant phages with longer inserts were obtained, the longest of which is approximately 850 bp. Nucleotide sequencing of this insert indicated that it is solely composed of the same tandem 60-nucleotide repeat unit. Similarly other longer cDNA inserts obtained by 3b cDNA probing of the library are also only composed of the tandem repeating unit.

Restriction enzyme digestion of the isolated 850-bp insert with *Sma*I (CCCGGG) completely reduces it to 60-bp fragments (data not shown) thus indicating that *Sma*I sites appear at 60-bp intervals.

A comparison of the repeat units found in the 3b cDNA with those present in the 850-bp insert (Fig. 1 A, B and C) at 70°C for 60 min and the membranes were then rehybridized with a different probe under similar conditions.

Construction of eukaryotic expression vector coding for H23 Ag

The *Xmn*I – *Eco*RI genomic fragment (see Fig. 6) was inserted into the eukaryotic expression vector pCL642 (this vector will be described in detail in a separate publication). Briefly, pCL642 is composed of the promoter region (1.4 kb) isolated from the mouse housekeeping gene coding for 3-hydroxy-3-methylglutaryl-coenzyme-A reductase. The promoter is followed by the untranslated first exon and intron (0.7 kb and 3.5 kb) derived also from this reductase gene. The *Xmn*I site of the *Xmn*I – *Eco*RI genomic fragment coding for the tumor antigen was blunt-end-ligated to an *Eco*RV site located in a polylinker immediately downstream to the reductase intron. The *Eco*RI site was ligated to the *Kpn*I site of the polylinker via an *Eco*RI – *Kpn*I adaptor. A 123-bp fragment containing the SV40 poly(A) signal sequence is situated immediately 3′ to the polylinker. The construct pCL642/*Xmn*I – *Eco*RI (10 μg) was contransfected with 1 μg pAG60(G418R) plasmid [22] into either MM5tC3H cells (from American Tissue Culture Collection) or FR3T3 ras-1 cells [23] using a modification [24] of the calcium phosphate precipitation method [25]. Cells were selected for G418 resistance (Geniticin 500 μg/ml) and loci were picked and subcultured.

For the detection of tumor antigen, the transfected cells were grown on coverslips and immunohistochemically stained with H23 mAbs. Control cells were either transfected with the pAG60 plasmid alone or with an irrelevant gene.

Nucleotide sequencing

Sequencing was accomplished using the dideoxynucleotide chain-termination method [26]. Restriction fragments of the cDNA inserts were subcloned into M13 and both strands were sequenced. The ssDNA was primed with either the M13 universal primer or synthetic oligonucleotides prepared according to known sequences. The analysis of the sequence was performed using the Beckman MicroGenie program.

Radioactive labelling of DNA probes

Double-stranded DNA probes were radioactively labelled with [α-^{32}P]dCTP either by nick translation or random oligonucleotide multipriming using commercially available kits (BRL, USA, and Amersham, England, respectively). All DNA probes used here were purified inserts that were isolated by agarose gel electrophoresis. Single-stranded oligonucleotides were 5′-end labelled by incubating with [γ-^{32}P]ATP and polynucleotide kinase. All labelled probes were purified from non-incorporated nucleotide by passage through Sephadex G-100 columns.

Oligonucleotide synthesis

Oligonucleotides were prepared at the Macromolecule Synthesis Service Unit (Department of Organic Chemistry, Weizman Institute of Science) by Dr Ora Goldberg using an Automated Applied Biosystems synthesizer. Following synthesis, the oligonucleotides were electrophoretically purified on acrylamide/urea gels.

Immunoblotting

Protein samples denatured by boiling in SDS buffer containing mercaptoethanol were analyzed on 3 – 15% linear gradient SDS/acrylamide gels as previously described [27]. Following electrophoresis, the gel was equilibrated in transfer buffer (Tris/glycine) and electrotransferred for 3 h at 1 A to nitrocellulose filters. The filters were blocked in NaCl/P$_i$ (150 mM NaCl, 15 mM sodium phosphate, pH 7.0) containing 5% skimmed milk (Blotto) followed by incubation with antibody in Blotto. The filters were washed in NaCl/P$_i$ and then reacted with ^{125}I-protein A (Amersham, England) dissolved in Blotto.

Monoclonal antibodies

Monoclonal antibodies (mAbs) were prepared against particulate antigens released into the medium by T47D breast carcinoma cells, using established procedures. The monoclonal antibodies obtained were screened against paraffin-embedded sections of benign and malignant breast tissue with the immunoperoxidase-staining technique and one of the mAbs designated H23 [2] was selected and used in this study.

RESULTS

Sequence of cDNA coding for epitope recognized by H23 monoclonal antibodies

The gt11 cDNA expression libraries prepared with mRNA isolated from either T47D or MCF7 [17], both human breast carcinoma cell lines, were probed with the monoclonal antibody H23. Libraries obtained by priming poly(A)-rich RNA with oligo(dT), as well as with random nucleotide oligomers, were investigated. Recombinant clones immunoreactive with H23, were obtained at a frequency of approximately 1 in 2000 in the amplified libraries and the cDNA inserts of all clones analyzed revealed a size of approximately 220 – 240 bp. Both random oligomer-primed as well as oligo(dT)-primed libraries from the MCF7 and T47D cell lines revealed similar-sized inserts.

Nucleotide sequencing of one such representative cDNA insert, termed 3b, indicates that it is (G + C)-rich with strand preference for the G or C nucleotides. Inspection of the 225-bp sequence shows that it is composed of a 60-nucleotide tandem repeat unit which is remarkably conserved with only very few substitutions occurring between the different units (Fig. 1 A).

Longer cDNA inserts contain the same 60-bp tandem repeat unit

In order to obtain longer cDNA clones, the 3b cDNA insert was used to reprobe the library by DNA/DNA hybridization. Several recombinant phages with longer inserts were obtained, the longest of which is approximately 850 bp. Nucleotide sequencing of this insert indicated that it is solely composed of the same tandem 60-nucleotide repeat unit. Similarly other longer cDNA inserts obtained by 3b cDNA probing of the library are also only composed of the tandem repeating unit.

Restriction enzyme digestion of the isolated 850-bp insert with *Sma*I (CCCGGG) completely reduces it to 60-bp fragments (data not shown) thus indicating that *Sma*I sites appear at 60-bp intervals.

A comparison of the repeat units found in the 3b cDNA with those present in the 850-bp insert (Fig. 1 A, B and C)

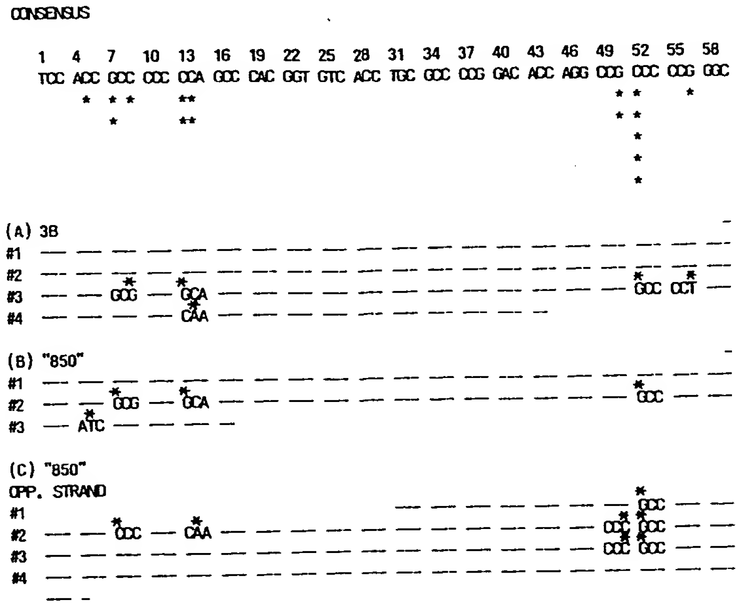


Fig. 1. *Nucleotide sequence of cDNAs that code for the epitope recognized by H23 mAbs and contain 60-bp tandem repeat units*. The gt11 cDNA expression library was screened with H23 mAbs as described in Methods and the cDNA insert (indicated as 3B but referred to in the text as 3b) obtained from a positive purified recombinant phage was subcloned in M13 vectors in both orientations and sequenced (A). The 3b cDNA insert was purified, nick-translated and used to reprobe the library under stringent hybridization conditions as described in Methods. The longest cDNA insert ('850', i.e. 850 bp) thus obtained was subcloned in M13 and both strands were partially sequenced (B and C). Only the C-rich strand is presented. The consensus sequence of the 60-bp repeat unit is shown at the top of the figure. Nucleotides in the repeat units identical to this sequence are indicated with dashes whilst substitutions are shown by an asterisk

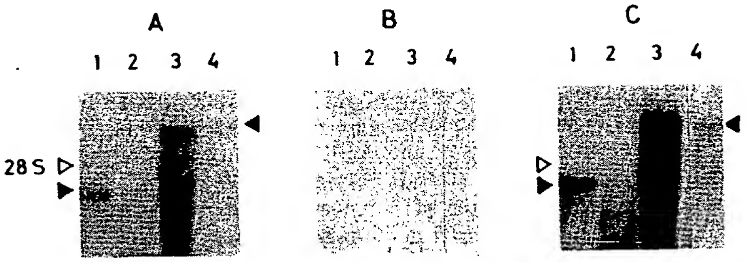


Fig. 2. *Northern blot analysis of human breast tumor RNA samples with 3b cDNA probe and synthetic complementary oligonucleotides derived from the repeating unit*. RNA was extracted from human breast tumor tissue (lanes 1 and 3) or adjacent 'normal' breast tissue (lanes 2 and 4) from two separate individuals (lanes 1 + 2 and 3 + 4) and analyzed by glyoxal agarose gel electrophoresis followed by Northern blotting to nylon membranes and hybridization with (A) the 3b cDNA probe, (B) the C-rich oligonucleotide 5' AGCCCACGGTGTCACCTCGGCCCCGGACA 3' identical to nucleotides 15–43 of the concensus sequence presented in Fig. 1 A and (C) the complementary G-rich oligonucleotide 5' TGTCCGGGGCCGAGGTGACACCGTGGGCT 3'. The probe used in A was radioactively labelled by nick translation whilst those used in B and C were end-labelled by polynucleotide kinase and [γ-^{32}P]ATP as described in Methods. The blots were stringently washed and autoradiographed at −70°C. The full arrow to the left of the figure indicates the 3.6-kb hybridizing mRNA species whilst that on the right points to the 6.0-kb mRNA species detected in the other sample. The open arrow indicates the position of 28S rRNA

demonstrates remarkable conservation with very few nucleotide substitutions occurring between the repeats.

Coding strand of the tandem repeat unit

To determine the coding strand of the tandem repeat unit, Northern blots were probed with synthetic oligonucleotides complementary to either strand of the repeat unit. Probing a Northern blot containing RNA isolated from human breast samples (both tumor and adjacent 'normal' tissues) shows that the G-rich synthetic oligonucleotide hybridizes to mRNA species (Fig. 2). An identical hybridization pattern was observed when 3b cDNA was used to probe the same blot (Fig. 2). The breast tumor tissue has in the one case a hybridizing mRNA species of 6.5 kb whilst the second sample shows a single band at 3.6 kb. The corresponding RNA samples from adjacent 'normal' tissue are identically sized but much reduced in amount. In contrast to the above results, no hybridization at all is seen with the second complementary C-rich oligonucleotide (Fig. 2). These findings confirm that RNA species containing multiple 60-nucleotide tandem repeats are *bonafide* transcripts. Moreover the orientation of transcription is demonstrated and the C-rich strand of the cDNA insert is the coding strand.

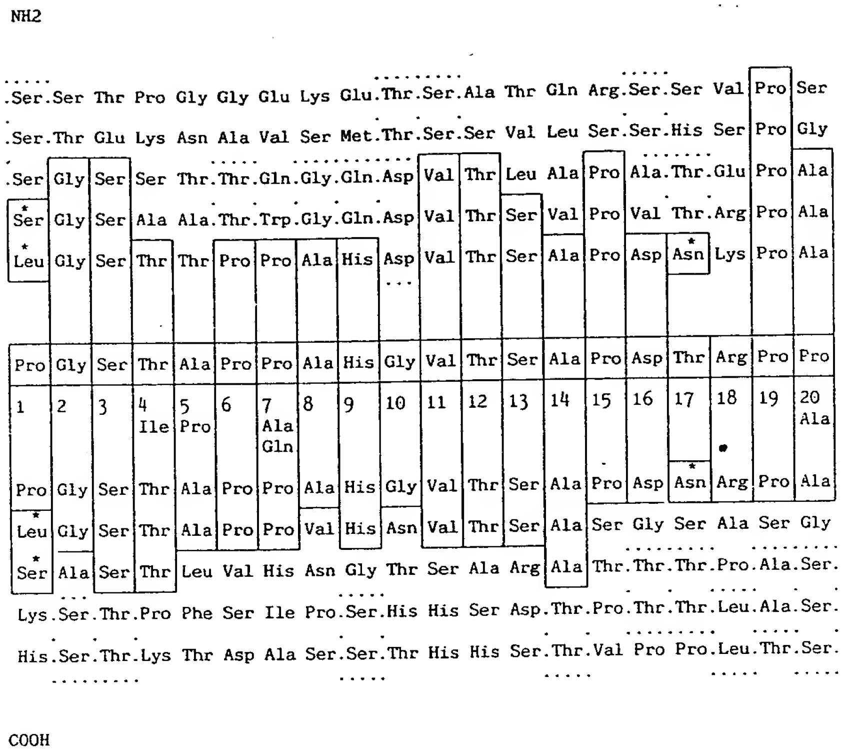

NH2
.Ser.Ser Thr Pro Gly Gly Glu Lys Glu.Thr.Ser.Ala Thr Gln Arg.Ser.Ser Val Pro Ser
.Ser.Thr Glu Lys Asn Ala Val Ser Met.Thr.Ser.Ser Val Leu Ser.Ser.His Ser Pro Gly
.Ser Gly Ser Ser Thr.Thr.Gln.Gly.Gln.Asp Val Thr Leu Ala Pro Ala.Thr.Glu Pro Ala
*Ser Gly Ser Ala Ala.Thr.Trp.Gly.Gln.Asp Val Thr Ser Val Pro Val Thr.Arg Pro Ala
*Leu Gly Ser Thr Thr Pro Pro Ala His Asp Val Thr Ser Ala Pro Asp *Asn Lys Pro Ala
Pro Gly Ser Thr Ala Pro Pro Ala His Gly Val Thr Ser Ala Pro Asp Thr Arg Pro Pro
1 2 3 4 Ile 5 Pro 6 7 Ala Gln 8 9 10 11 12 13 14 15 16 17 18 19 20 Ala
Pro Gly Ser Thr Ala Pro Pro Ala His Gly Val Thr Ser Ala Pro Asp *Asn Arg Pro Ala
*Leu Gly Ser Thr Ala Pro Pro Val His Asn Val Thr Ser Ala Ser Gly Ser Ala Ser Gly
*Ser Ala Ser Thr Leu Val His Asn Gly Thr Ser Ala Arg Ala Thr.Thr.Thr.Pro.Ala.Ser.
Lys.Ser.Thr.Pro Phe Ser Ile Pro.Ser.His His Ser Asp.Thr.Pro.Thr.Thr.Leu.Ala.Ser.
His.Ser.Thr.Lys Thr Asp Ala Ser.Ser.Thr His His Ser.Thr.Val Pro Pro.Leu.Thr.Ser.
COOH

Fig. 3. *Comparative analysis of the flanking amino acid sequences with the 20-amino-acid repeat motif.* The amino acid sequence of the repeat motif is presented in the central boxed region and numbered from 1 to 20. The alternative amino acids that occur due to variations in the consensus sequence are indicated below the numbers. The 100 amino acids flanking the repeat motif on the amino and carboxyl terminals are shown (NH_2 and COOH, respectively). Flanking amino acids that are identical with the repeat motif are boxed in by the full line, whereas the flanking amino acids that appear in the same position every 20 amino acids are boxed in by a series of dots. Amino acids that vary from the repeat motif and appear at the same positions on either side of the repeat motif are indicated by * and are boxed in

Comparative analysis of the flanking amino acid sequences with the 20-amino-acid repeat motif

The determined coding strand of the 60-bp cDNA could be translated in all three reading frames. As almost full-length cDNAs coding for the H23 Ag have recently been isolated [16a], the correct reading frame of the repeat motif could be readily identified (Fig. 3). The high level of nucleotide conservation amongst the various repeat units is reflected in the repeat-unit amino acid sequences (Fig. 3). The studies reported here (see also below) show that the tumor antigen has the unusual structure of highly conserved repeat units that compose at least 50% of the protein molecule. It is therefore of considerable interest to compare the similarity of flanking non-repeat amino acid sequences with the 20-amino-acid repeat motif itself.

Several possibilities may be envisaged. (a) An abrupt break may occur in the continuity of the repeat motif and no similarity exist between the flanking amino acid sequences and the repeat units. (b) Some degree of similarity may exist between the flanking amino acid sequences and the repeat motif that declines with increasing distance away from the repeat array. Or (c) the flanking amino acid sequences may retain similarity only with specific amino acids or regions of the consensus repeat motif.

The comparative analysis (Fig. 3) shows that indeed similarity exists between the flanking amino acid sequences and the repeat motif itself. However, this similarity is confined to specific subregions such as the Val-Thr-Ser and Gly-Ser peptides at residues 11 – 13 and 2 – 3, respectively, and occurs in the flanking amino acid sequences on both sides of the repeat motif. On the other hand, certain amino acid residues are conserved asymmetrically, i.e. either upstream or downstream to the repeat motif. Significant conservation in the amino-terminal flanking sequences occurs with the proline residues (15 and 19) and the alanine residue (20), whereas threonine (4) and alanine (14) are appreciably conserved only in the carboxyl-terminal flanking sequences. The conservation of the proline residue (19) in the upstream flanking sequences is particularly remarkable as it is located in the same position for 82 amino acids upstream to the repeat motif. Of further note are the amino acids that diverge from the repeat motif: the asparagine residue that replaces threonine (17) does so on both sides flanking the repeat motif. Furthermore, proline (1) is replaced by leucine and followed 20 amino acids later by serine; it is indeed striking that identical changes occur both in the upstream and downstream flanking sequences.

At the present time, the significance of repeat-motif amino acid conservation, as well as the identical amino acid changes occurring on both sides of the repeat, is not known. They may impose certain structural constraints on the protein molecule or/and be related to a function involving specific subregions of the repeat motif.

Expression of tumor antigen mRNA in primary human tissues

The *in vivo* system studies were extended and we investigated the presence of mRNA species hybridizing with the 3b cDNA probe in a variety of benign and malignant human tissues by RNA dot blotting and Northern blot analysis. The initial dot-blot screening demonstrates very significant levels of hybridizing mRNA species in total RNA prepared from a

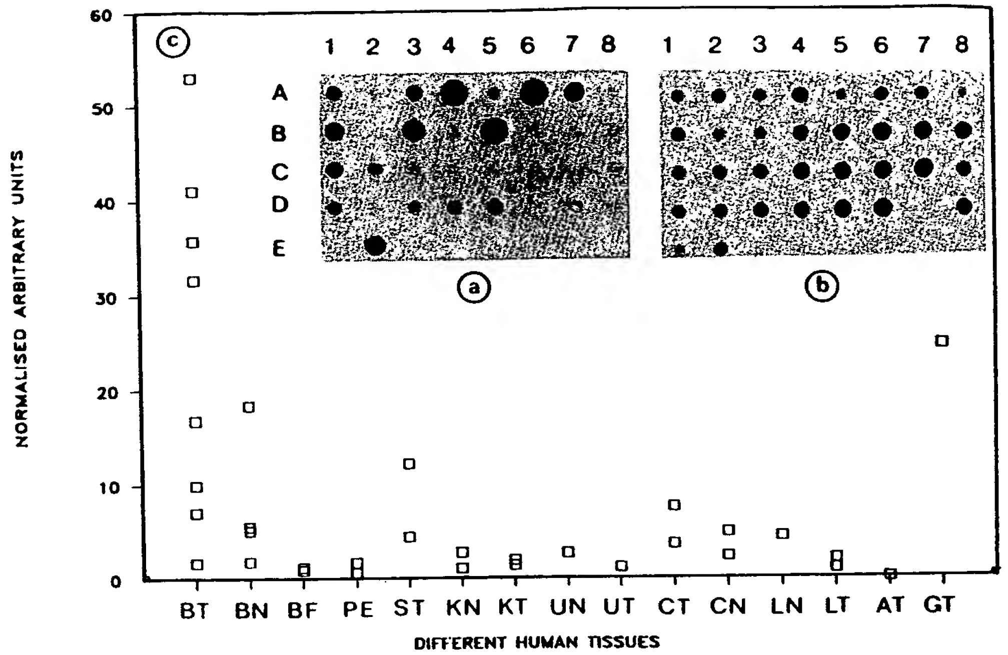


Fig. 4. *Levels of RNA species in human tissues hybridizing with the 3b cDNA probe*. Total RNA from different human tissues was dot-blotted and probed with (a) the 3b cDNA insert and (b) a cDNA insert (unpublished results) coding for part of human 18S ribosomal RNA. The key for the dot blots is as follows: (A1 – 8) BT2, BT4, BT5, BT6, BN7, BT7, BN9, BN10; (B1 – 8) BT10, BN12, BT12, BF13, BT15, BF16, PE1, PE2; (C1 – 8) ST1, ST2, KN1, KN2, KT2, UT1, UN1; (D1 – 8) CT1, CN1, CT2, LN1, LT1, blank, LT2; (E1 – 2) AT, GT1. Abbreviations used are for RNA extracted from: BT = breast adenocarcinoma, ST = gastric carcinoma, KT = hypernephroma, UT = transitional cell carcinoma, CT = colon adenocarcinoma, LT = lung tumor, AT = pheochromocytoma and GT = ovarian carcinoma. The corresponding 'N' samples (for example BN) represent RNA isolated from 'normal' tissue adjacent to the tumor. The same numbers indicate preparations from the same patient. The BF samples are breast fibroadenomas. PE samples are from pleural effusion metastatic cells of patients with advanced breast cancer. (c) The dot blots were scanned by laser densitometry using an LKB laser densitometer and the absorbance values obtained with the blot probed with the 3b cDNA probe were divided by the levels observed following 18S cDNA probing. This procedure resulted in a normalized arbitrary unit corresponding to each sample which is presented on the ordinate of the figure. Total RNA extraction, dot blotting, hybridization and washing conditions were as described in Materials and Methods. The blots were exposed to Agfa Curix X-ray films at –70°C with a Curix special intensifying screen

number of breast carcinomas (Fig. 4A). Several of these samples contain large quantities of mRNA capable of hybridizing with the 3b cDNA probe and a quantitative analysis demonstrates high levels of mRNA hybridization to the 3b cDNA probe (Fig. 4C). Significantly lower levels of hybridization are observed in RNA isolated from non-malignant breast tissue adjacent to the biopsied tumor sample. For example, Fig. 4A shows that RNA isolated from tissue adjacent to tumor in samples BN7 (B = breast, N = normal), BN10 and BN12 (dot-blot positions A5, A8 and B2, respectively) demonstrate hybridizing values of 5.5, 5.1 and 1.8 (normalized probe-specific hybridization) whereas 3b cDNA hybridization to RNA extracted from the corresponding tumor samples BT7 (B = breast, T = tumor), BT10 and BT12 (dot-blot positions A6, B1 and B3, respectively) shows considerably higher values of 53.0, 16.8 and 41.1 respectively). Two breast fibroadenomas (dot-blot positions B4 and B6) contain very low levels of hybridizing RNA. Of all benign breast tissues analyzed to date, only one sample that was pathologically classified as nonmalignant (BN9-A7 on the dot blot) contains significant levels of 3b cDNA hybridizing RNA. Interestingly, this sample was obtained from the second breast of a breast cancer patient who had undergone mastectomy several years earlier. In contrast, mostly low, albeit detectable, levels of hybridization to the 3b cDNA probe are present in RNA extracted from stomach, colon and lung adenocarcinomas, as well as hypernephroma. Extremely low levels are seen in RNA isolated from a bladder carcinoma and undetectable levels of hybridization occurred with RNA from an adrenal pheochromocytoma, as well as in RNA extracted from chronic lymphocytic leukemic cells or from a brain neuroblastoma sample (data not shown).

Analyses of human tumor RNA species by Northern blotting

In order to determine by an independent method the validity of the dot-blot analysis, the human-tissue RNAs were analyzed by probing Northern blots with the 3b cDNA probe (Fig. 5). Differences in both the sizes of the hybridizing mRNA species as well as in the relative levels are immediately evident. The relative levels obtained in the Northern blot analysis correlate well with those seen in the initial dot-blot screening. The most intensive hybridization is observed with RNA extracted from the breast tumor BT15 which yields a prominent RNA band located at the 6.5-kb position along with a significantly weaker band at approximately 3.6 kb (Fig. 5B, lane 7). A densitometric analysis indicates that hybridization in this breast-tumor RNA sample is 30 – 40-fold higher than that observed in other RNA samples analyzed on the same Northern blot. Much lower levels of hybridization of a 3.6-kb species are seen in RNA isolated from the two pleural effusions (lanes 5 and 9) which were revealed by longer exposure of the autoradiograph (data not shown).

RNA extracted from the breast fibroadenoma (lane 6, Fig. 5B) demonstrates only very low levels of hybridization with a 3.0-kb RNA species, whereas barely detectable levels

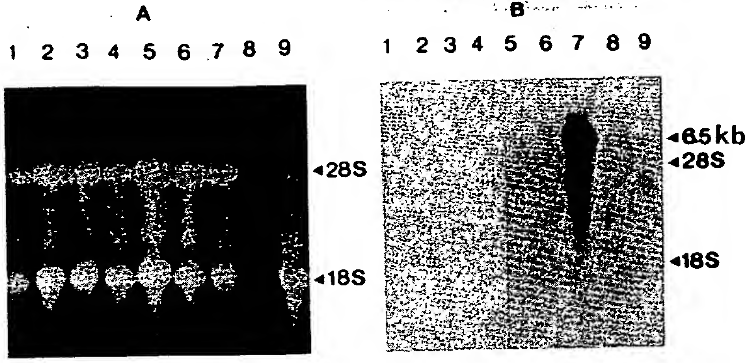


Fig. 5. *Northern blot analysis of RNA species hybridizing with 3b cDNA*. RNA samples isolated from a human hypernephroma KT1 (lane 1), adjacent kidney non-malignant tissue KN1 (lane 2), thyroid nodular goiter (lane 3), thyroid follicular adenoma and lymphocytic thyroiditis (lane 4), cells from the pleural effusions of two patients with advanced breast carcinoma PE1 and PE2 (lanes 5, 9), breast fibroadenoma BF16 (lane 6), breast adenocarcinoma BT15 (lane 7) and brain neuroblastoma (lane 8) were analyzed by agarose gel electrophoresis and ethidium bromide staining (A) followed by Northern blotting and probing with the 3b cDNA insert. (B) The autoradiograph of 1-day exposure is presented. Total RNA was extracted from tissue samples and analyzed by Northern blotting and probing with the 3b cDNA insert as described in Materials and Methods. All washings were performed under stringent conditions (0.2 × NaCl/Cit, twice for 30 min at 60 °C)

of hybridization at the 3.6-kb position are seen in one of the thyroid samples (lane 4, longer exposure, data not shown). Hybridization is not detected with RNA from a neuroblastoma, one thyroid sample and non-malignant tissue of the hypernephroma (lanes 8, 3 and 2, respectively).

The gene hybridizing with the repeat unit is polymorphic and is a VNTR gene: verification with unique non-repeat genomic sequences

The presence of the 60-nucleotide tandem repeat unit in the cDNAs analyzed indicates that the gene coding for this protein probably also contains a variable number of tandem repeats and thus belongs to the class known as VNTR genes. In order to demonstrate the polymorphism occurring in such a gene, a Southern blot comprising *Eco*RI and *Eco*RI/*Pst*I double-digested DNA was prepared from a number of human tissue samples isolated from different individuals. Hybridization with either the 3b cDNA insert or with the larger 850-bp cDNA insert (previously described above) shows marked gene polymorphism with at least 11 different alleles evident in the 9 samples studied (Fig. 6). Although the allelic patterns are similar on the *Eco*RI or double-digested *Eco*RI/*Pst*I DNA samples, the sizes of the different alleles following the double digestion are significantly smaller, thus increasing their electrophoretic resolution (Fig. 6).

From the Southern blot and cDNA nucleotide sequencing data presented, it is concluded that (a) the different alleles result from differences in the number of repeat units, (b) the *Eco*RI and *Pst*I sites are situated outside the tandem repeat unit and (c) the *Pst*I sites are closer to the borders of the tandem repeat units than are the *Eco*RI sites.

Polymorphism of this gene has also recently been described by two other groups [13–16]. However, the only probe used in the reported studies has been the 60-bp repeat unit [13–16]. Conclusive evidence that the gene is in fact an expressed VNTR gene requires probing of both Northern and Southern blots also with unique non-repeat sequences that are linked to the repeat array.

We further verified the VNTR nature of the gene by reprobing the same Southern blot with a non-repeat DNA fragment excised from the cloned 7.5-kb *Eco*RI–*Eco*RI gene fragment, isolated from a genomic library by probing with the cDNA 60-bp repeat unit [34]. This fragment (a *Sma*I–*Pst*I fragment, see Fig. 6), is approximately 1 kb and is situated 5′ to the array of tandem repeat units. It should thus hybridise with a single identically sized DNA band in all samples that have been *Eco*RI/*Pst*I double-digested. On the other hand, hybridization of this same fragment with an *Eco*RI-digested DNA should yield an identical hybridization pattern to that seen following hybridization with the repeat unit. These predictions were confirmed by the results obtained. Hybridization of the *Eco*RI/*Pst*I-digested samples with the 1-kb non-repeat fragment reveals a 3.5-kb band in all samples investigated (Fig. 6B, *Eco*RI + *Pst*I). This band is absent when the blot is probed with the repeat unit (compare Fig. 6A and B *Eco*RI + *Pst*I digest). The lightly labelled additional bands designated by asterisks in Fig. 6B are the remnants of the first hybridization with the repeat unit (compare with Fig. 6A, *Eco*RI + *Pst*I). Furthermore, hybridization of the *Eco*RI digest with the non-repeat fragment or with the repeat unit are identical, as predicted (compare Fig. 6A and B *Eco*RI).

As expected, the larger allele that contains more repeat units than the smaller allele shows a stronger signal following hybridization with the repeat unit probe (see Fig. 6, lane 2 for example).

The above data present evidence that the gene coding for the tumor-associated antigen is indeed a VNTR gene.

The different alleles are codominantly transcribed into corresponding mRNA species

We had previously seen significant heterogeneity in mRNA species that hybridize with 3b cDNA expressed in tumor samples isolated from different individuals. Two, and less often, only one hybridizing RNA band(s) are observed in any individual sample. As the gene itself is highly polymorphic, we investigated whether a correlation exists between the different allelic forms and the number and sizes of hybridizing RNAs expressed (Fig. 7).

Although, as noted above, 3b-hybridizing mRNAs are highly over-expressed in most malignant breast tissues, RNA isolated from other epithelial tumors also demonstrate hybridizing mRNA species albeit at lower levels. In order to establish the scope of this possible allele/mRNA correlation, investigations were performed both on non-breast and breast tumor

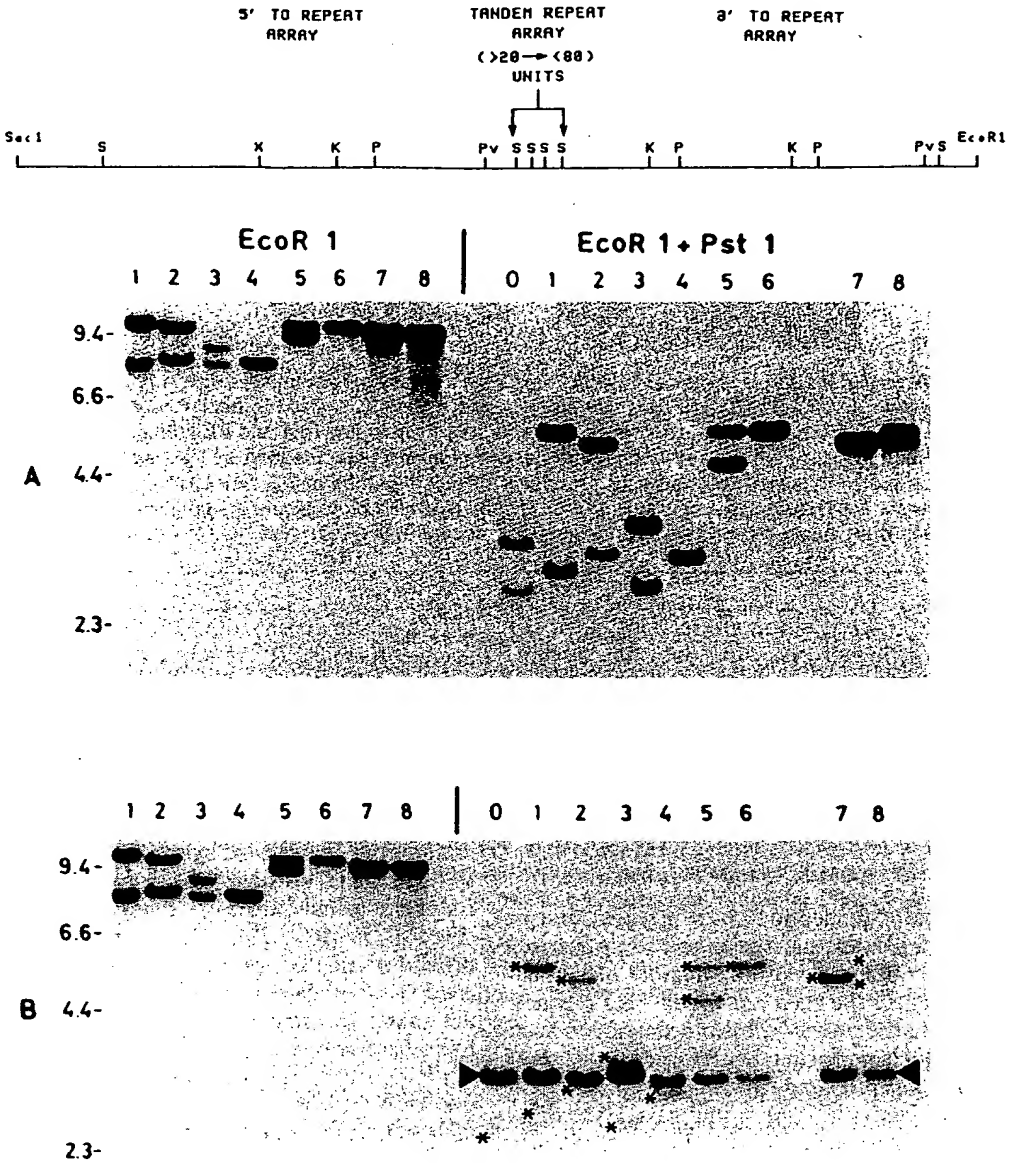


Fig. 6. *Hybridization of Southern blots with the repeating unit demonstrates a highly polymorphic gene.* High-molecular-mass genomic DNA was extracted from the following human organs that had malignant tumors: stomach (lane 0), ovary (lanes 1, 2), lung (lane 3), breast (lanes 4 and 6), colon (lane 5) and thyroid (lanes 7 and 8). The DNA was restricted with *Eco*RI alone or doubly digested with *Eco*RI and *Pst*I. (A) 10 μg was electrophoresed on agarose gels, Southern blotted and probed with radioactively labelled 850-bp cDNA. (B) Following this hybridization, the blot was rehybridized with a 1-kb non-repeat fragment of the gene (restriction map of gene, top panel, *Sma*I – *Pst*I fragment 5′ to the repeat array). The restriction enzymes *Kpn*I, *Pst*, *Pvu*II, *Sma*I and *Xmn*I are represented by K, P, Pv, S and X respectively. The blots were stringently washed and autoradiographed at −70°C. The bands labelled in B with the asterisk are the remaining signals of those seen in the previous hybridization with the repeat unit, whereas the specifically labelled band is shown by the full arrow (B, *Eco*RI + *Pst*I). The numbers to the left of the figure indicate size (kb) of markers

samples. Two breast tumor samples that express the lowest levels of 3b-hybridizing mRNA out of all malignant breast tissues analyzed were selected for comparison. In this regard it should be emphasized that the investigation was performed on nucleic acids isolated from primary human tissues rather than from cell culture lines. The conclusions of these experiments are therefore relevant to the *in-vivo* situation. In 10 out of 10 primary human tumor samples investigated, full concordance is demonstrated between the number and sizes of alleles with the corresponding hybridizing mRNA species (Fig. 7A and B).

The different allelic forms probably vary due to a difference in the number of tandem repeats. We thus investigated whether the corresponding mRNA species expressed in the same individual demonstrate an identical size difference. As the homozygotic breast tumor samples correspondingly express one mRNA species, they were not included in this analysis. The results shown in Fig. 7 indicate that, within the accuracy possible for DNA fragment and mRNA species size determination, the allelic size difference for the heterozygotic samples is equal to the difference in size of the two mRNA species.

It is interesting to note that the mRNA species correlating with the larger allele gives a less intense hybridization signal than the smaller mRNA species (see Fig. 7, lanes 2, 3 and 8). We do not know whether this is a consequence of reduced

omic DNA
reast (lanes
nd *Pst*I. (A)
lowing this
tI fragment
y. The blots
f those seen
- *Pst*I). The

sed in the
rence. As
dingly ex-
this analy-
the accu-
pecies size
erozygotic
o mRNA

orrelating
ion signal
, 3 and 8).
f reduced

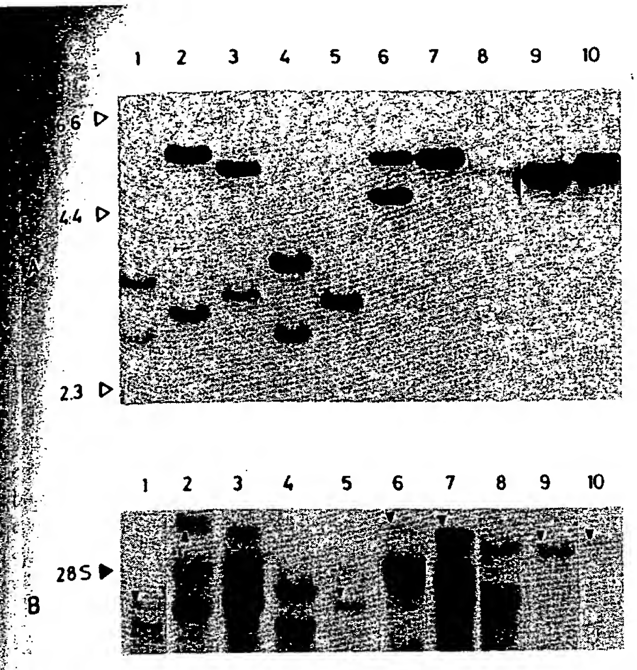


Fig. 7. *Correlation of Southern and Northern blots containing DNA and RNA isolated from the same human tissue sample following hybridization with the repeating unit.* DNA (A) or total RNA (B) was isolated from the following human tissues that had malignant tumors: stomach (lane 1), ovary (lanes 2 and 3), lung (lane 4), breast (lanes 5 and 7) and thyroid (lanes 8–10). The DNA after double digestion with *Eco*RI and *Pst*I or total RNA samples were electrophoresed on agarose gels and Southern (A) or Northern (B) blotted followed by hybridization with radioactively labelled 3b cDNA probe. The blots were stringently washed and autoradiographed at −70 °C. All samples in A were run simultaneously on the same gel but lane 8 was exposed for a longer time as less DNA was available for analysis. On the Northern blot, samples 2–7 were run simultaneously on the same gel but lane 6 was exposed for a longer time as there was significantly less mRNA expression in this sample. Samples 1 and samples 8–10 were run on two separate gels. The positions of the hybridizing mRNA species are indicated in B by the upward or downward facing full arrows. Note that in lanes 2–7 (and especially in lane 6) some nonspecific hybridization with 28S rRNA has occurred

transcription of the larger allele, reduced stability of the larger mRNA species or other mechanisms.

In order to characterize further the correlation of allelic forms with the different mRNA species, both a Southern and Northern blot were rehybridized with the 1-kb non-repeat genomic fragment described above (Fig. 8A and B). As expected, probing the Northern blot with either the 3b cDNA tandem repeat units or with the 1-kb non-repeat fragment (Fig. 8B, lanes 1 and 2 respectively) reveals identical hybridizing mRNA species. On the other hand, reprobing the Southern blot with the 1-kb non-repeat fragment demonstrates only one band in contrast to the two allelic forms seen following probing with the repeat units (Fig. 8A, lanes 2 and 1 respectively).

Expression of H23 Ag in cells stably transfected with the H23 Ag gene

By probing Northern and Southern blots with both unique genomic sequences and the 60-bp repeat unit, we demonstrated the expression of a VNTR gene that codes for the H23 Ag. These critical experiments hinge on the physical linkage, in the genomic fragment isolated, of unique non-repeat DNA sequences with the tandem repeat array.

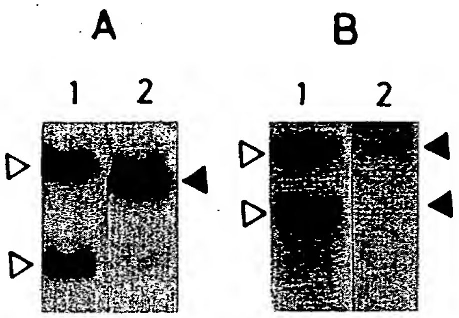


Fig. 8. *Probing Southern and Northern blots with the repeating-unit cDNA probe and a non-repeat genomic fragment.* High-molecular-mass DNA (A) or total RNA (B) was isolated from human lung tissue. The DNA, following *Eco*RI plus *Pst*I double digestion, and total RNA were electrophoresed on agarose gels and Southern (A) or Northern blotted (B). The blots were hybridized with the 3b repeating-unit cDNA probe (lane 1) or, after stripping, with the 1-kb non-repeat fragment of the gene (lane 2, see text). The blots were stringently washed and autoradiographed at −70°C. The full arrows indicate bands hybridizing with the 1-kb non-repeat genomic fragment whilst the open arrows show bands hybridizing to the 3b repeat-unit cDNA probe. The efficiency of stripping the Northern blot following the first hybridization with 3b cDNA (B, lane 1) was evaluated by blot autoradiography prior to the second hybridization: no signal at all was seen. The bands appearing in B, lane 2, are thus *bona fide* signals

In order to confirm this linkage, mouse or rat cells were transfected with the isolated genomic fragment and then analyzed for H23 Ag synthesis. We had previously determined by cDNA and genomic sequencing (unpublished results) that an *Xmn*I site is located 35 nucleotides upstream to the putative ATG initiation codon of the H23 Ag gene. The *Xmn*I – *Eco*RI gene fragment (see Fig. 6) was therefore isolated and inserted into a eukaryotic expression vector downstream to the promoter of a housekeeping gene, 3-hydroxy-3-methylglutaryl-coenzyme-A reductase. In order to obtain stable transfectants, the H23 Ag gene construct, pCL642/*Xmn-Eco*, was cotransfected into mouse mammary tumor cells MM5, with a plasmid coding for resistance to the antibiotic neomycin. Similar transfections were conducted with c-Ha-ras-transformed rat fibroblasts. Neither of these cell lines expressed any human epithelial tumor antigen detectable with H23 mAb. As a control, MM5 cells were separately stably transfected with a pCL642 construct containing an irrelevant gene. (Details on the pCL642 eukaryotic expression vector are to be presented in a separate publication.)

Both the MM5 and rat fibroblast stable transfectants were grown on coverslips and immunohistochemically stained with H23 mAb (Fig. 9). Whereas no staining is observed in control MM5 and rat fibroblasts transfected with the non-relevant gene (Fig. 9A′ and B′), stable transfectants harboring the pCL642/*Xmn-Eco* construct demonstrate intense staining, readily detected with the H23 mAb (Fig. 9A and B). Staining is mainly cytoplasmic and is undetectable within the cell nucleus.

Western blot analyses of cell proteins from the pCL642/*Xmn-Eco* transfection demonstrate high-molecular-mass proteins (five bands ranging from 70 kDa to > 200 kDa) that are immunoreactive with H23 mAb (data not shown). These protein species are likely to represent H23 Ag glycosylated to varying degrees, thereby producing heterogenously sized immunoreactive products. Cell extracts from cells transfected with the non-relevant gene show no immunoreactive bands on the Western blot analysis.

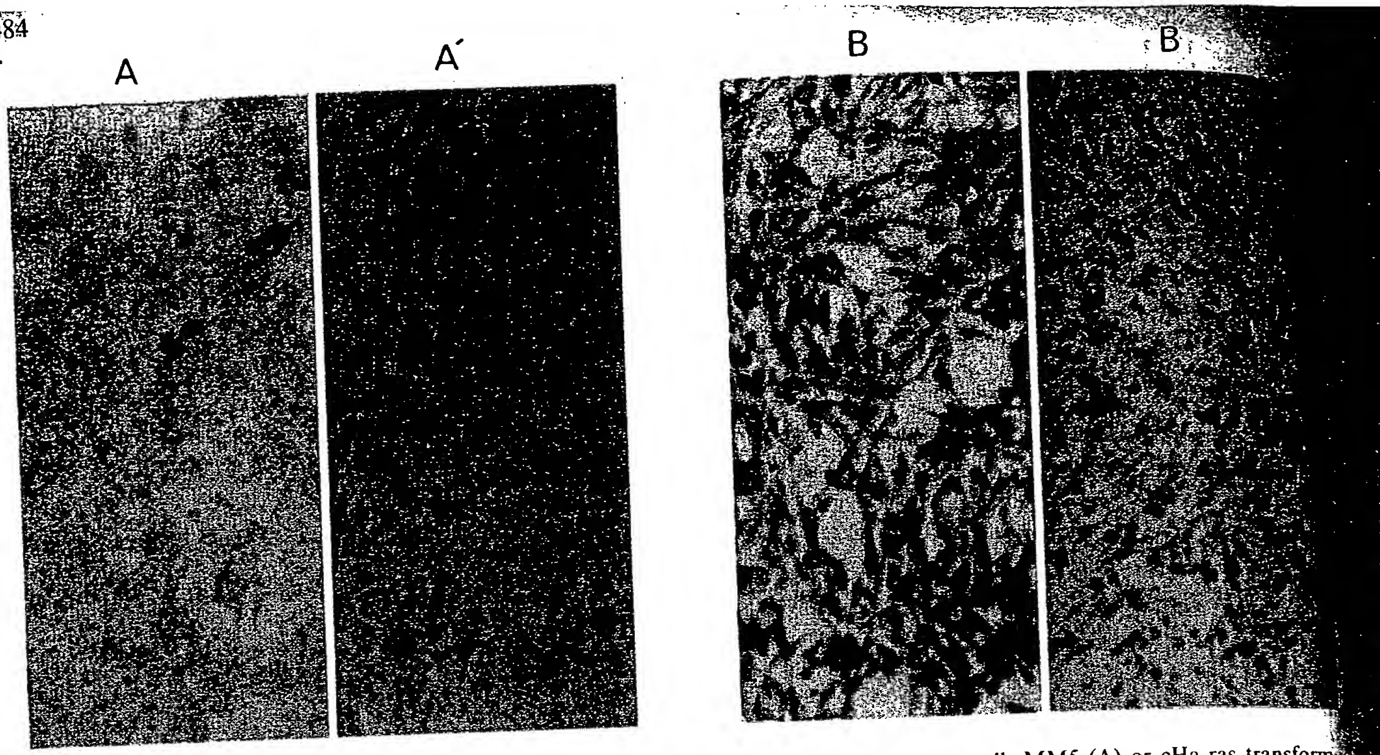


Fig. 9. *Expression of H23 Ag in cells transfected with the H23 Ag gene.* Mouse mammary tumor cells MM5 (A) or cHa-ras transformed fibroblasts (B) were transfected with the H23 Ag gene inserted into an expression vector, as described in Methods, grown on coverslips and immunohistochemically stained with H23 mAbs. Controls were MM5 cells (A′) or cHa-ras transformed fibroblasts (B′) transfected with a non-relevant gene and stained with H23 mAb. Intense cytoplasmic and membrane-bound staining is observed in the H23 Ag gene-transfected cells (A and B)

Over-expression of the H23 Ag in primary human breast tumor tissue: Western blot analysis

Having established (a) that in primary human tissues the gene polymorphism directly correlates with the mRNA species expressed and (b) that the mRNA coding for the antigen is over-expressed in breast tumor tissue, we next investigated the expression of antigen at the protein level.

A preliminary investigation was conducted on the human breast cell line T47D, which expresses large amounts of tumor antigen that are readily detectable by H23 mAbs. These cells were analysed at the gene, mRNA and protein levels.

A Southern blot of an *Eco*RI/*Pst*I digest shows two allelic forms at 5.5 and 3.1 kb (Fig. 10A, lane s). The Northern blot analysis correspondingly demonstrates two mRNA species (6.5 and 4.1 kb) that hybridize with the repeat-unit cDNA probe (data not shown). The protein products of these mRNA species were analyzed by immunoblotting which shows two products migrating on SDS-denaturing gels with molecular masses in the region of 250 – 450 kDa (Fig. 10A, lane w). No bands are observed when the immunoblot was probed with a non-specific monoclonal antibody under identical conditions. The alleles of the T47D gene are thus transcribed into mRNA species that are subsequently translated into distinct high-molecular-mass protein products that correlate with the respective mRNA and allelic sizes.

In order to relate the above findings to an *in-vivo* system, these studies were extended, as with the RNA and DNA analyses, to primary human tissues. Extracts of human tissue samples were run on SDS-denaturing gels and the separated protein species immunoblotted and probed with H23 mAbs. Analyses were performed on malignant breast tumor tissue samples together with an extract from non-malignant breast tissue adjacent to the biopsied tumor sample.

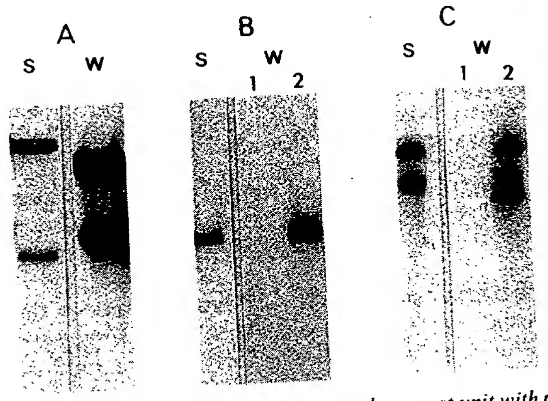


Fig. 10. *Correlation of alleles hybridizing to the repeat unit with protein products detected by H23 monoclonal antibodies and H23 Ag overexpression in breast cancer tissue.* DNA, double-digested with *Eco*RI and *Pst*I, or RNA isolated from T47D breast carcinoma cells were analyzed by agarose gel electrophoresis and Southern (A, s) or Northern blotted (not shown). The blots were hybridized to the 3b cDNA repeating unit probe, stringently washed and autoradiographed. For immunoblotting (A, B and C, w), samples were boiled for 3 min in SDS/mercaptoethanol sample buffer, and electrophoresed on a 4 – 15% SDS gradient gel, followed by electro-transfer to a nitrocellulose membrane as in Methods, and reacted with H23 mAb. Sample A, lane w is the medium of T47D cells precipitated with 50% ammonium sulfate; samples B and C are the protein extracts from breast cancer tissue (lane 2) and the adjacent non-malignant breast tissue (lane 1). DNA extracted from the same samples (B and C, lanes s) was restricted with *Eco*RI alone, Southern blotted and probed with 3b cDNA

Probing the immunoblots with H23 mAbs demonstrates marked over-expression of the tumor antigen in the malignant breast tissue samples (Fig. 10B and C, lane 2). The non-maligbreast tissues, that were adjacent to their malignant counterparts, show significantly lower immune reactivity with the H23 mAbs (Fig. 10B and C, lane 1).

As previously shown for T47D, the polymorphism of the H23-immunoreactive protein species seen in the primary human breast tissue samples correlates with the different allelic forms observed in Southern blots probed with the 60-bp repeat unit (Fig. 10B and C, lane s).

DISCUSSION

The results presented here show that a highly polymorphic gene contains a 60-bp tandem repeat array and codes for an epithelial tumor antigen that is over-expressed in human breast cancer. The H23 monoclonal antibody recognizes an epitope contained within the 20-amino-acid repeat motif encoded by the 60-bp cDNA and detects intracytoplasmic antigen in 91% of malignant breast tumors [2]. An almost identical 60-bp cDNA insert has been isolated by two other groups [13–16] using monoclonal antibodies (DF3, HMFG-1, HMFG-2 and SM3) that also recognize high-molecular-mass mucin-like glycoproteins aberrantly expressed in breast cancer tissue. It seems likely that different post-translational modifications occur within the 20-amino-acid repeat motif, encoded by the 60-bp cDNA, thus explaining, in part, the varying specificities of the different mAbs for normal and malignant breast tissue.

The gene coding for tumor antigen

As previously reported [13–16], the gene coding for the tumor antigen contains a variable number of tandem repeats and is highly polymorphic. We have extended this finding and probed restricted genomic DNA samples with unique non-repeat sequences isolated from a genomic fragment that contains the tandem repeat array. This analysis demonstrates that, external to the repeat array, the gene does not exhibit any heterogeneity, thereby indicating that the genetic polymorphism is solely due to varying numbers of the 60-bp tandem repeats. It is also demonstrated here that, besides expression of the 60-bp repeat units, unique non-repeat genomic sequences are expressed into mRNA and translated into protein.

The physical linkage of unique non-repeat sequences with the expressed 60-bp tandem repeat array was further confirmed by transfection experiments. The isolated gene fragment, from which the unique repeat sequences are derived, was transfected into mouse and rat cells that do not normally express any tumor antigen detectable with H23 mAb. The transfectants thus obtained synthesize human tumor antigen that is readily detected by H23 mAb. Furthermore, these transfection studies provide strong evidence that the isolated cDNA and gene fragment are indeed *bona fide* sequences that code for the human epithelial tumor antigen.

Correlation of alleles with expressed mRNA species and protein products: studies with primary human tissues

In a recent study involving only material derived from cells grown in culture [13], the gene polymorphism was found to correlate with both the mRNA species and protein forms detected. The different protein species observed in human urine by immunoblot analysis [16] also correlate with the various alleles. To our knowledge, there have been no reports demonstrating a concordance between the various alleles, mRNA species and protein forms in primary human tissues. We show here that in primary human tumors full concordance exists between the alleles and the transcribed mRNA species. This is demonstrated for nucleic acids extracted from breast, ovary, lung, stomach and thyroid tissues. Furthermore, it is shown that the allelic and mRNA size differences are equivalent in every sample of primary human tissue analyzed. These studies indicate that the heterogeneity in mRNA species is also solely due to the number of tandem repeats that they contain.

The correlative study of alleles and mRNA species in the same samples allows us to determine that approximately 1.9 kb in any individual mRNA species is represented by non-repeat sequences. The coding capacity of the tandem array is thus probably greater than 50% of the total protein, even in the smallest mRNA observed, and could code for more than 65% in the larger mRNA species.

Analyses of RNA samples from primary benign and malignant tumors demonstrate undetectable levels of hybridization in tissues of nonepithelial origin, whereas several non-mammary epithelial adenocarcinoma tumors display low levels of hybridization. However, RNA extracted from three ovarian carcinomas shows significant levels of hybridization with the 3b cDNA (an example of the intensity of hybridization is shown in Fig. 4A, dot-blot position E2). A question of obvious interest is whether this expression is due to the endocrine nature of these tissues.

The highest levels of hybridizing mRNA species are detected in malignant breast tumors. Non-malignant 'normal' tissue adjacent to the breast tumor samples, as well as non-malignant breast fibroadenomas, display much lower hybridization levels. The increased expression of the mRNA species hybridizing with the 3b cDNA probe thus strongly correlates with the malignant phenotype of the breast tissue. Although the mechanisms involved in the increased expression are not known, they may be related to the de-differentiated state of malignant tissue.

Since H23 mAbs detect an intracellular antigen primarily in breast tumor sections [2], the detection of hybridizing RNA from non-breast tumors with the 3b cDNA probe, albeit at low levels, is surprising. We have recently isolated unique-sequence cDNA that account for almost full-length cDNA of the tumor antigen [16a]. As several different alternatively spliced cDNAs were characterized, it is possible that the loss of epitope recognition by H23 mAb may be due to alternative splicing of the mRNA species in non-breast tissues. Other possibilities to be considered are different translational frames or simply a question of sensitivity of the immunohistochemical staining technique.

The expression of the gene coding for the tumor antigen was also investigated at the protein level. This study shows that in primary human breast tumor tissue the polymorphism detected in the gene and mRNA species correlates with the protein products detected by immunoblotting. Moreover, it is quite obvious that the malignant breast tissue contains significantly higher levels of tumor antigen than adjacent normal breast tissue.

We and others [12] have described a 68-kDa protein species that can be precipitated with mAbs directed against this epithelial tumor antigen. The 68-kDa protein is not detected, however, using the conditions of the immunoblot technique described here. It may represent a partially glycosylated protein or alternatively a breakdown product (induced proteolytically or otherwise), that contains a discrete number of repeat motifs. These possibilities are presently being investigated.

The 20-amino-acid repeat motif: comparison with flanking amino acids

The epithelial tumor antigen is composed of 20-amino-acid repeats that make up more than 50% of the total protein. This unusual structure of highly conserved repeat motifs has recently been documented for porcine submaxillary gland mucin [28], human intestinal mucin [29], a cell-surface antigen expressed by murine hemopoietic progenitor cells [30], human apolipoprotein (a) [31], apo-polysialoglycoprotein of rainbow trout eggs [32] and a repetitive protein from *Xenopus laevis* skin [33]. It is interesting to note that repeat elements from several of these proteins are also rich in serine and threonine residues. The function of the 20-amino-acid repeat motif, as of the complete epithelial tumor antigen itself, is unknown. It is striking, however, that specific amino acids and subregions of the repeat element are conserved in the flanking regions on both sides of the repeat array. Furthermore, in some cases, identical amino acids replace a repeat motif amino acid in the same position on both sides of the repeat array. Although we do not understand the significance of this gradual decline in similarity between flanking amino acids and the repeat motif, it may indicate that a specific function is related to a certain amino acid sequence of the repeat motif.

Mouse cell transfectants synthesize the human epithelial tumor antigen: possible insights into tumor antigen function

Mouse cells transfected with the isolated gene coding for the human epithelial tumor antigen synthesize protein readily detected with H23 mAb. The location of synthesized antigen is primarily cytoplasmic, although we cannot rule out the possibility that it may be bound to the endoplasmic reticulum and/or plasma membrane. Recent analyses of full-length cDNAs [16a] show that differential splicing events occur 5′ and 3′ to the tandem repeat array giving mRNAs that will produce several forms of the antigen that localize to different cellular compartments.

Using the transfected cells as a model system, we are now in a position to ask questions regarding the function of this epithelial tumor antigen: does it change the growth characteristics, morphology or/and tumorgenicity of the transfected cells? Preliminary results indicate that expression of the tumor antigen indeed changes the cell growth potential; these and other possible functions are presently being investigated.

This work was supported by the Simko Chair for Breast Cancer Research, Frederico Fund for Tel Aviv University, grants from *Center National Récherche Scientifique*, Mr Toby Green (London), Maurice and Frida Eskanasy Cancer Research Fund and the Israel Cancer Association (DHW). DHW, MH and IT were the recipients of European Molecular Biology Organization short-term and long-term (MH) Fellowships; DHW is the recipient of a Koret Foundation Fellowship, California. We thank Professor Pierre Chambon for continued support and fruitful discussions.

REFERENCES

1. Keydar, I., Chen, L., Karby, S., Weiss, F. R., Delarea, J., Radu, M., Chaitchik, S. & Brenner, H. J. (1979) *Eur. J. Cancer 15*, 659–670.
2. Keydar, R., Chou, C. S., Hareuveni, M., Tsarfaty, I., Sahar, E., Seltzer, G., Chaitchik, S. & Hizi, A. (1989) *Proc. Natl Acad. Sci. USA 86*, 1362–1366.
3. Abe, M. & Kufe, D. (1986) *J. Cell. Physiol. 126*, 126–136.
4. Bramwell, M. E., Bhavanandan, V. P., Wiseman, G. & Ha H. (1983) *Br. J. Cancer 48*, 177–183.
5. Burchell, J. M., Durbin, H. & Taylor-Papadimitriou, J. (1983) *Immunol. 131*, 508–513.
6. Ceriani, R. L., Peterson, J. A. & Blank, E. W. (1984) *Cancer R 44*, 3033–3039.
7. Hilkens, J., Buijs, F., Hilgers, J., Hagemann, Ph., Calafat, Sonnenberg, A. & Van der Valk, M. (1984) *Int. J. Cancer* 197–206.
8. Johnson, V. G., Schlom, J., Paterson, A. J., Bennett, J., Magnan J. L. & Colcher, D. (1986) *Cancer Res. 46*, 850–857.
9. Lan, M. S., Finn, O. J., Fernsten, P. D. & Metzgar, R. S. (198 *Cancer Res. 45*, 305–310.
10. Magnani, J. L., Steplewski, Z., Koprowski, H. & Ginsburg, (1983) *Cancer Res. 43*, 5489–5492.
11. Price, M. R., Edwards, S., Robins, R. A., Hilgers, J., Hilkens J. & Baldwin, R. (1986) *Eur. J. Can. Clin. Oncol. 22*, 115–11
12. Burchell, J., Gendler, S., Taylor-Papadimitriou, J., Girling, A Lewis, A., Mills, R. & Lamport, D. (1987) *Cancer Res. 4* 5476–5482.
13. Siddiqui, J., Abe, M., Hayes, D., Shani, E., Yunis, E. & Kufe, D (1988) *Proc. Natl Acad. Sci. USA 85*, 2320–2323.
14. Gendler, S., Taylor-Papadimitriou, J., Duhig, T., Rothbard, J. & Burchell, J. (1988) *J. Biol. Chem. 263*, 12820–12823.
15. Gendler, S. J., Burchell, J. M., Duhig, T., Lamport, D., White R., Parker, M. & Taylor-Papadimitriou, J. (1987) *Proc. Natl Acad. Sci. USA 84*, 6060–6064.
16. Swallow, D. M., Gendler, S., Griffiths, B., Corney, G., Taylor-Papadimitriou, J. & Bramwell, M. E. (1987) *Nature 328*, 82–84.

16a. Wreschner, D. H., Hareuveni, M., Tsarfaty, I., Smorodinsky, N., Horev, J., Zaretsky, J., Kotkes, P., Weiss, M., Lathe, R., Dion, A. & Keydar, I. (1990) *Eur. J. Biochem. 189*, 463–473.

17. Petkovitch, M., Brand, N. J., Krust, A. & Chambon, P, (1987) *Nature 330*, 445–450.
18. Walter, P., Green, S., Green, G., Krust, A., Bornert, J. M., Jeltsch, J. M., Staub, A., Jensen, E., Scrace, G., Waterfield, M. & Chambon, P. (1985) *Proc. Natl Acad. Sci. USA 82*, 7889–7893.
19. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol. 113*, 237–251.
20. Wreschner, D. H. & Rechavi, G. (1988) *Eur. J. Biochem. 172*, 333–344.
21. Aviv, H. & Leder, P. (1972) *Proc. Natl Acad. Sci. USA 69*, 1408–1412.
22. Colbere-Garapin, F., Horodniceanu, F., Kourilsky, P. & Garapin, A. C. (1981) *J. Mol. Biol. 150*, 1–14.
23. Matriceau, L. M., Glaichenhaus, N., Gesnel, M. C. & Breathnach, R. (1985) *EMBO J. 4*, 1435–1440.
24. Wigler, S. F., Pellicer, A., Silverstein, S. & Axel, R. (1978) *Cell 14*, 725–731.
25. Graham, F. & Van der Eb, A. (1973) *Virology 52*, 456–457.
26. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl Acad. Sci. USA 74*, 5463–5467.
27. Laemmli, U. K. (1970) *Nature 227*, 680–685.
28. Timpte, C. S., Eckhardt, A. E., Abernethy, J. L. & Hill, R. L. (1988) *J. Biol. Chem. 263*, 1081–1088.
29. Gun, J. R., Byrd, J. C., Hicks, J. W., Toribara, J. W., Lamport, D. T. A. & Kim, Y. S. (1989) *J. Biol. Chem. 264*, 6480–6487.
30. Dougherty, G. J., Kay, R. J. & Humphries, R. K. (1989) *J. Biol. Chem. 264*, 6509–6514.
31. McLean, J. W., Tomlinson, J. E., Kuang, W. J., Eaton, D. L., Chen, E. Y., Fless, G. M., Scanu, A. M. & Lawan, R. M. (1987) *Nature 330*, 132–137.
32. Sorimachi, M., Emori, Y., Kawasaki, H., Kitajima, D., Inoue, S., Suzuki, K. & Inoue, Y. (1988) *J. Biol. Chem. 263*, 17678–17684.
33. Hoffman, W. (1988) *J. Biol. Chem. 263*, 7686–7690.
34. Tsarfaty, I., Hareuveni, M., Horev, J., Zaretsky, J., Weiss, M., Jeltsch, J. M., Garnier, J. M., Lathe, R., Keydar, I. & Wreschner, D. H. (1990) *Gene*, in the press.

NCBI PubMed National Library of Medicine NLM

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Bc

Search PubMed for Go Clear

☑ Limits Preview/Index History Clipboard Details

About Entrez

Display Abstract Show: 20 Sort Send to Text

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby

Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

Privacy Policy

□ 1: J Parasitol. 2003 Apr;89(2):381-4. Related Articles, Lin

Upregulation of cardiac cell plasma membrane calcium pump in a canine model of Chagas disease.

Barr SC, Pannabecker TL, Gilmour RF Jr, Chandler JS.

Department of Clinical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, New York 14853, USA. scb6@cornell.edu

We have previously demonstrated that cardiac myocytes isolated from the hearts of adult dogs develop rapid repetitive cytosolic Ca2+ transients, membrane depolarization, and cell contraction by mobilization of sarcoplasmic reticulum Ca2+ stores when exposed to a soluble factor from th trypomastigotes of Trypanosoma cruzi. These findings led us to investigate the regulatory mechanisms of cytosolic Ca2+ in cardiac tissues from dogs chronically infected with T. cruzi. Expression of the plasma membrane calcium pump (PMCA) RNA and protein was determined by Northern and Western blotting, respectively, followed by densitometric analyses. A 642-b PMCA 1b complementary DNA probe derived from canine epicardial tissue hybridized to 2 major transcripts (7.3 and 5.3 kb) in canine epicardium. Expression of the dominant transcript (7.3 kb) was 77% greater in cardiac tissues obtained from dogs with chronic T. cruzi infection (140 days after inoculation) in comparison with constitutive expression levels in normal dog Monoclonal antibody 5F10, known to recognize all isoforms of the PMCA, was used to detect expression of the PMCA protein in epicardial tissue. Expression of a 142-kDa protein was increased by 58% in the cardiac tissues of infected dogs when compared with those from uninfected dogs. To establish a link between the upregulation of PMCA in dogs chronically infected with Chagas disease and the ventricular-based arrhythmias and myocardial failure that occur during this stage of disease both in dogs and humans, further study will be required.

PMID: 12760659 [PubMed - indexed for MEDLINE]

Display Abstract Show: 20 Sort Send to Text

Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions

James U. Bowie,* John F. Reidhaar-Olson, Wendell A. Lim, Robert T. Sauer

An amino acid sequence encodes a message that determines the shape and function of a protein. This message is highly degenerate in that many different sequences can code for proteins with essentially the same structure and activity. Comparison of different sequences with similar messages can reveal key features of the code and improve understanding of how a protein folds and how it performs its function.

The genome is manifest largely in the set of proteins that it encodes. It is the ability of these proteins to fold into unique three-dimensional structures that allows them to function and carry out the instructions of the genome. Thus, comprehending the rules that relate amino acid sequence to structure is fundamental to an understanding of biological processes. Because an amino acid sequence contains all of the information necessary to determine the structure of a protein (*1*), it should be possible to predict structure from sequence, and subsequently to infer detailed aspects of function from the structure. However, both problems are extremely complex, and it seems unlikely that either will be solved in an exact manner in the near future. It may be possible to obtain approximate solutions by using experimental data to simplify the problem. In this article, we describe how an analysis of allowed amino acid substitutions in proteins can be used to reduce the complexity of sequences and reveal important aspects of structure and function.

Methods for Studying Tolerance to Sequence Variation

There are two main approaches to studying the tolerance of an amino acid sequence to change. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection. This method has been extremely powerful for proteins such as the globins or cytochromes, for which sequences from many different species are known (2–7). The second approach uses genetic methods to introduce amino acid changes at specific positions in a cloned gene and uses selections or screens to identify functional sequences. This approach has been used to great advantage for proteins that can be expressed in bacteria or yeast, where the appropriate genetic manipulations are possible (*3, 8–11*). The end results of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding or function. If a particular property of a side chain, such as charge or size, is important at a given position, only side chains that have the required property will be allowed. Conversely, if the chemical identity of the side chain is unimportant, then many different substitutions will be permitted.

Studies in which these methods were used have revealed that proteins are surprisingly tolerant of amino acid substitutions (*2–4, 11*). For example, in studying the effects of approximately 1500 single amino acid substitutions at 142 positions in *lac* repressor, Miller and co-workers found that about one-half of all substitutions were phenotypically silent (*11*). At some positions, many different, nonconservative substitutions were allowed. Such residue positions play little or no role in structure and function. At other positions, no substitutions or only conservative substitutions were allowed. These residues are the most important for *lac* repressor activity.

What roles do invariant and conserved side chains play in proteins? Residues that are directly involved in protein functions such as binding or catalysis will certainly be among the most conserved. For example, replacing the Asp in the catalytic triad of trypsin with Asn results in a 10^4-fold reduction in activity (*12*). A similar loss of activity occurs in λ repressor when a DNA binding residue is changed from Asn to Asp (*13*). To carry out their function, however, these catalytic residues and binding residues must be precisely oriented in three dimensions. Consequently, mutations in residues that are required for structure formation or stability can also have dramatic effects on activity (*10, 14–16*). Hence, many of the residues that are conserved in sets of related sequences play structural roles.

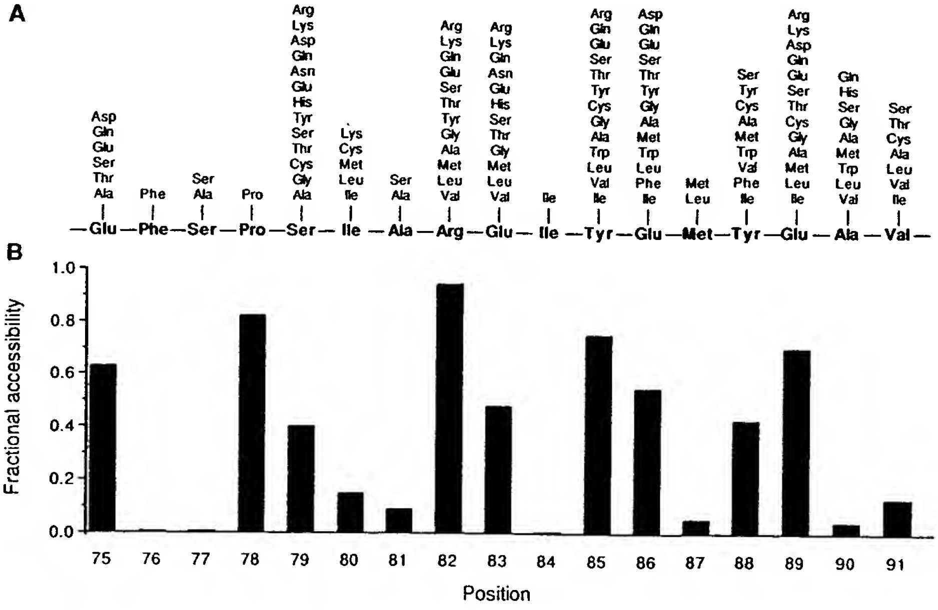
Substitutions at Surface and Buried Positions

In their initial comparisons of the globin sequences, Perutz and co-workers found that most buried residues require nonpolar side chains, whereas few features of surface side chains are generally conserved (*6*). Similar results have been seen for a number of protein families (*2, 4, 5, 7, 17, 18*). An example of the sequence tolerance at surface versus buried sites can be seen in Fig. 1, which shows the allowed substitutions in λ repressor at residue positions that are near the dimer interface but distant from the DNA binding surface of the ... were identified by a functiona

The authors are in the Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.

... Biochemistry and the Molecular

Fig. 1. (A) Amino acid substitutions allowed in a region of λ repressor. The wild-type sequence is shown along the center line. The allowed substitutions shown above each position were identified by randomly mutating one to three codons at a time by using a cassette method and applying a functional selection (9). (B) The fractional solvent accessibility (42) of the wild-type side chain in the protein dimer (43) relative to the same atoms in an Ala-X-Ala model tripeptide.



selection after cassette mutagenesis. A histogram of side chain solvent accessibility in the crystal structure of the dimer is also shown in Fig. 1. At six positions, only the wild-type residue or relatively conservative substitutions are allowed. Five of these positions are buried in the protein. In contrast, most of the highly exposed positions tolerate a wide range of chemically different side chains, including hydrophilic and hydrophobic residues. Hence, it seems that most of the structural information in this region of the protein is carried by the residues that are solvent inaccessible.

Constraints on Core Sequences

Because core residue positions appear to be extremely important for protein folding or stability, we must understand the factors that dictate whether a given core sequence will be acceptable. In general, only hydrophobic or neutral residues are tolerated at buried sites in proteins, undoubtedly because of the large favorable contribution of the hydrophobic effect to protein stability (19). For example, Fig. 2 shows the results of genetic studies used to investigate the substitutions allowed at residue positions that form the hydrophobic core of the NH_2-terminal domain of λ repressor (20). The acceptable core sequences are composed almost exclusively of Ala, Cys, Thr, Val, Ile, Leu, Met, and Phe. The acceptability of many different residues at each core position presumably reflects the fact that the hydrophobic effect, unlike hydrogen bonding, does not depend on specific residue pairings. Although it is possible to imagine a hypothetical core structure that is stabilized exclusively by residues forming hydrogen bonds and salt bridges, such a core would probably be difficult to construct because hydrogen bonds require pairing of donors and acceptors in an exact geometry. Thus the repertoire of possible structures that use a polar core would probably be extremely limited (21). Polar and charged residues are occasionally found in the cores of proteins, but only at positions where their hydrogen bonding needs can be satisfied (22).

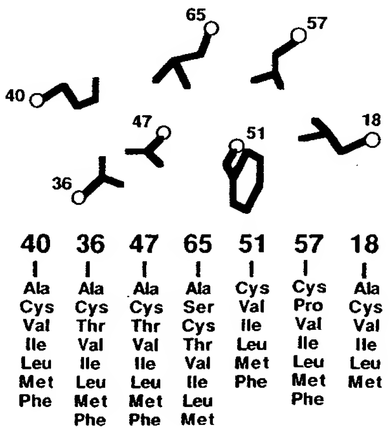
The cores of most proteins are quite closely packed (23), but some volume changes are acceptable. In λ repressor, the overall core volume of acceptable sequences can vary by about 10%. Changes at individual sites, however, can be considerably larger. For example, as shown in Fig. 2, both Phe and Ala are allowed at the same core position in the appropriate sequence contexts. Large volume changes

phylogenetic studies, where it has been noted that the size decreases and increases at interacting residues are not necessarily related in a simple complementary fashion (5, 7, 17). Rather, local volume changes are accommodated by conformational changes in nearby side chains and by a variety of backbone movements.

The Informational Importance of the Core

With occasional exceptions, the core must remain hydrophobic and maintain a reasonable packing density. However, since the core is composed of side chains that can assume only a limited number of conformations (24), efficient packing must be maintained without steric clashes. How important are hydrophobicity, volume, and steric complementarity in determining whether a given sequence can form an acceptable core? Each factor is essential in a physical sense, as a stable core is probably unable to tolerate unsatisfied hydrogen bonding groups, large holes, or steric overlaps (25). However, in an informational sense, these factors are not equivalent. For example, in experiments in which three core residues of λ repressor were mutated simultaneously, volume was a relatively unimportant informational constraint because three-quarters of all possible combinations of the 20 naturally occurring amino acids had volumes within the range tolerated in the core, and yet most of these sequences were unacceptable (20). In contrast, of the sequences that contained only

Fig. 2. Amino acid substitutions allowed in the core of λ repressor. The wild-type side chains are shown pictorially in the approximate orientation seen in the crystal structure (43). The lists of allowed substitutions at each position are shown below the wild-type side chains. These substitutions were identified by randomly mutating one to four residues at a time by using a cassette method and applying a functional selection (20). Not all substitutions are allowed in every sequence back-


65
57
40
47
51
18
36
40 Ala Cys Val Ile Leu Met Phe
36 Ala Cys Thr Val Ile Leu Met Phe
47 Ala Cys Thr Val Ile Leu Met Phe
65 Ala Ser Cys Thr Val Ile Leu Met
51 Cys Val Ile Leu Met Phe
57 Cys Pro Val Ile Leu Met Phe
18 Ala Cys Val Ile Leu Met

the appropriate hydrophobic residues, a significant fraction were acceptable. Hence, the hydrophobicity of a sequence contains more information about its potential acceptability in the core than does the total side chain volume. Steric compatibility was intermediate between volume and hydrophobicity in informational importance.

The Informational Importance of Surface Sites

We have noted that many surface sites can tolerate a wide variety of side chains, including hydrophilic and hydrophobic residues. This result might be taken to indicate that surface positions contain little structural information. However, Bashford *et al.*, in an extensive analysis of globin sequences (*4*), found a strong bias against large hydrophobic residues at many surface positions. At one level, this may reflect constraints imposed by protein solubility, because large patches of hydrophobic surface residues would presumably lead to aggregation. At a more fundamental level, protein folding requires a partitioning between surface and buried positions. Consequently, to achieve a unique native state without significant competition from other conformations, it may be important that some sites have a decided preference for exterior rather than interior positions. As a result, many surface sites can accept hydrophobic residues individually, but the surface as a whole can probably tolerate only a moderate number of hydrophobic side chains.

Identification of Residue Roles from Sets of Sequences

Often, a protein of interest is a member of a family of related sequences. What can we infer from the pattern of allowed substitutions at positions in sets of aligned sequences generated by genetic or phylogenetic methods? Residue positions that can accept a number of different side chains, including charged and highly polar residues, are almost certain to be on the protein surface. Residue positions that remain hydrophobic, whether variable or not, are likely to be buried within the structure. In Fig. 3, those residue positions in λ repressor that can accept hydrophilic side chains are shown in orange and those that cannot accept hydrophilic side chains are shown in green. The obligate hydrophobic positions define the core of the structure, whereas positions that can accept hydrophilic side chains define the surface.

Functionally important residues should be conserved in sets of active sequences, but it is not possible to decide whether a side chain is functionally or structurally important just because it is invariant or conserved. To make this distinction requires an independent assay of protein folding. The ability of a mutant protein to maintain a stably folded structure can often be measured by biophysical techniques, by susceptibility to intracellular proteolysis (*26*), or by binding to antibodies specific for the native structure (*27, 28*). In the latter cases, it is possible to screen proteins in mutated clones for the ability to fold even if these proteins are inactive. Sets of sequences that allow formation of a stable structure can then be compared to the sets that allow both folding and function, with the active site or binding residues being those that are variable in the set of stable proteins but invariant in the set of functional proteins. The DNA-binding residues of Arc repressor were identified by this method (*8*). The receptor-binding residues of human growth hormone were also identified by comparing the stabilities and activities of a set of mutant sequences (*28*). However, in this case, the mutants were generated as hybrid sequences between growth hormone and related hormones with different binding specificities.

Implications for Structure Prediction

At present, the only reliable method for predicting a low-resolution tertiary structure of a new protein is by identifying sequence similarity to a protein whose structure is already known (*29, 30*). However, it is often difficult to align sequences as the level of sequence similarity decreases, and it is sometimes impossible to detect statistically significant sequence similarity between distantly related proteins. Because the number of known sequences is far greater than the number of known structures, it would be advantageous to increase the reach of the available structural information by improving methods for detecting distant sequence relations and for subsequently aligning these sequences based on structural principles. In a normal homology search, the sequence database is scanned with a single test sequence, and every residue must be weighted equally. However, some residues are more important than others and should be weighted accordingly. Moreover, certain regions of the protein are more likely to contain gaps than others. Both kinds of information can be obtained from sequence sets, and several techniques have

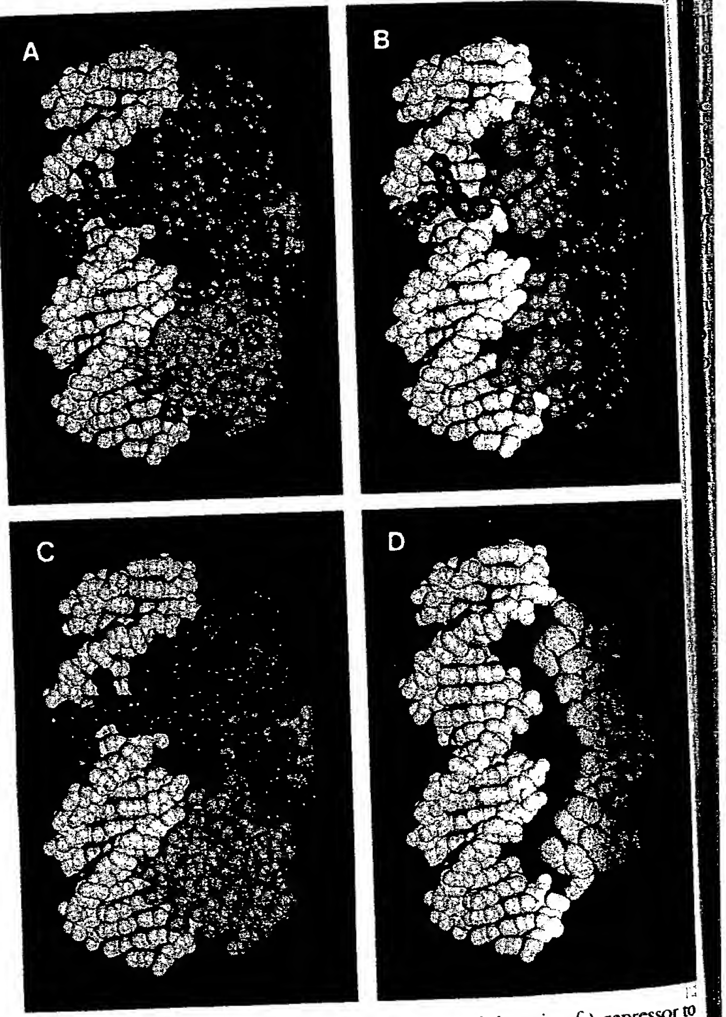


Fig. 3. Tolerance of positions in the NH$_2$-terminal domain of λ repressor to hydrophilic side chains. The complex (*43*) of the repressor dimer (blue) and operator DNA (white) is shown. In (**A**), positions that can tolerate hydrophilic side chains are shown in orange. The same side chains are shown in (**B**) without the remaining protein atoms. In (**C**), positions that require hydrophobic or neutral side chains are shown in green. These side chains are shown in (**D**) without the remaining protein atoms. About three-fourths of the 92 side chains in the NH$_2$-terminal domain are included in both (B) and (D). The remaining positions have not been tested. Data are from (*9, 14, 20, 27, 44*).

been used to combine such information into more appropriately weighted sequence searches and alignments (*31*). These methods were used to align the sequences of retroviral proteases with aspartic proteases, which in turn allowed construction of a three-dimensional model for the protease of human immunodeficiency virus type 1 (*29*). Comparison with the recently determined crystal structure of this protein revealed reasonable agreement in many areas of the predicted structure (*32*).

The structural information at most surface sites is highly degenerate. Except for functionally important residues, exterior positions seem to be important chiefly in maintaining a reasonably polar surface. The information contained in buried residues is also degenerate, the main requirement being that these residues remain hydrophobic. Thus, at its most basic level, the key structural message in an amino acid sequence may reside in its specific pattern of hydrophobic and hydrophilic residues. This is meant in an informational sense. Clearly, the precise structure and stability of a protein depends on a large number of detailed interactions. It is possible, however, that structural prediction at a more primitive level can be accomplished by concentrating on the most basic informational aspects of an amino acid sequence. For example, amphipathic patterns can be extracted from aligned sets of sequences and used, in some cases, to identify secondary structures.

If a region of secondary structure is packed against the hydrophobic core, a pattern of hydrophobic residues reflecting the periodicity of the secondary structure is expected (*33, 34*). These patterns can be obscured in individual sequences by hydrophobic residues on the protein surface. It is rare, however, for a surface position to remain hydrophobic over the course of evolution. Consequently, the amphipathic patterns expected for simple secondary structures can be much clearer in a set of related sequences (*6*). This principle is illustrated in Fig. 4, which shows helical hydrophobic moment plots of the Antennapedia homeodomain sequence (Fig. 4A) and for a composite sequence derived from a set of homologous homeodomain proteins (Fig. 4B) (*35*). The hydrophobic moment is a simple measure of the degree of amphipathic character of a sequence in a given secondary structure (*34*). The amphipathic character of the three α-helical regions in the Antennapedia protein (*36*) is clearly revealed only by the analysis of the combined set of homeodomain sequences. The secondary structure of Arc repressor, a small DNA-binding protein, was recently predicted by a similar method (*8*) and confirmed by nuclear magnetic resonance studies (*37*).

The specific pattern of hydrophobic and hydrophilic residues in an amino acid sequence must limit the number of different structures a given sequence can adopt and may indeed define its overall fold. If this is true, then the arrangement of hydrophobic and hydrophilic residues should be a characteristic feature of a particular fold. Sweet and Eisenberg have shown that the correlation of the pattern of hydrophobicity between two protein sequences is a good criterion of their structural relatedness (*38*). In addition, several studies indicate that patterns of obligatory hydrophobic positions identified from aligned sequences are distinctive features of sequences that adopt the same structure (*4, 29, 38, 39*). Thus, the order of hydrophobic and hydrophilic residues in a sequence may actually be sufficient information to determine the basic folding pattern of a protein sequence.

Although the pattern of sequence hydrophobicity may be a characteristic feature of a particular fold, it is not yet clear how such patterns could be used for prediction of structure de novo. It is important to understand how patterns in sequence space can be related to structures in conformation space. Lau and Dill have approached this problem by studying the properties of simple sequences composed only of H (hydrophobic) and P (polar) groups on two-dimensional lattices (*40*). An example of such a representation is shown in Fig. 5. Residues adjacent in the sequence must occupy adjacent squares on the lattice, and two residues cannot occupy the same space. Free energies of particular conformations are evaluated with a single term, an attraction of H groups. By considering chains of ten residues, an exhaustive conformational search for all 1024 possible sequences of H and P residues was possible. For longer sequences only a representative fraction of the allowed sequence or conformation space could be explored. The significant results were as follows: (i) not all sequences can fold into a "native" structure and only a few sequences form a unique native structure; (ii) the probability that a sequence will adopt a unique native structure increases with chain length; and (iii) the native states are compact, contain a hydrophobic core surrounded by polar residues, and contain significant secondary structure. Although the gap between these two-dimensional simulations and three-dimensional structures is large, the use of simple rules and sequence representations yields results similar to those expected for real proteins. Three-dimensional lattice methods are also beginning to be developed and evaluated (*41*).

Summary

There is more information in a set of related sequences than in a single sequence. A number of practical applications arise from an analysis of the tolerance of residue positions to change. First, such information permits the evaluation of a residue's importance to the function and stability of a protein. This ability to identify the essential elements of a protein sequence may improve our understanding of the determinants of protein folding and stability as well as protein function. Second, patterns of tolerance to amino acid substitutions of varying hydrophilicity can help to identify residues likely to be buried in a protein structure and those likely to occupy

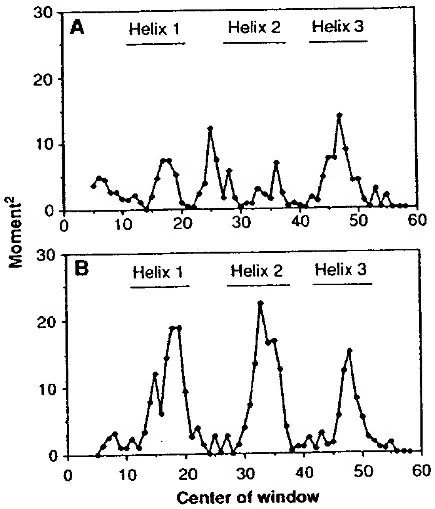


Fig. 4. Helical hydrophobic moments calculated by using (**A**) the Antennapedia homeodomain sequence or (**B**) a set of 39 aligned homeodomain sequences (*35*). The bars indicate the extent of the helical regions identified in nuclear magnetic resonance studies of the Antennapedia homeodomain (*36*). To determine hydrophobic moments, residues were assigned to one of three groups: H1 (high hydrophobicity = Trp, Ile, Phe, Leu, Met, Val, or Cys); H2 (medium hydrophobicity = Tyr, Pro, Ala, Thr, His, Gly, or Ser); and H3 (low hydrophobicity = Gln, Asn, Glu, Asp, Lys, or Arg). For the aligned homeodomain sequences, the residues at each position were sorted by their hydrophobicity by using the scale of Fauchere and Pliska (*45*). Arg and Lys were not counted unless no other residue was found at the position, because they contain long aliphatic side chains and can thereby substitute for nonpolar residues at some buried sites. To account for possible sequence errors and rare exceptions, the most hydrophilic residue allowed at each position was discarded unless it was observed twice. The second most hydrophilic residue was then chosen to represent the hydrophobicity of each position. An eight-residue window was used and the vectors projected radially every 100°. The vector magnitudes were assigned a value of 1, 0, or −1 for positions where the hydrophobicity group was H1, H2, or H3, respectively.

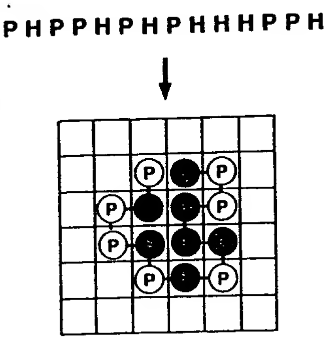


Fig. 5. A representation of one compact conformation for a particular sequence of H and P residues on a two-dimensional square lattice. [Adapted from (*40*), with permission of the American Chemical Society]

surface positions. The amphipathic patterns that emerge can be used to identify probable regions of secondary structure. Third, incorporating a knowledge of allowed substitutions can improve the ability to detect and align distantly related proteins because the essential residues can be given prominence in the alignment scoring.

As more sequences are determined, it becomes increasingly likely that a protein of interest is a member of a family of related sequences. If this is not the case, it is now possible to use genetic methods to generate lists of allowed amino acid substitutions. Consequently, at least in the short term, it may not be necessary to solve the folding problem for individual protein sequences. Instead, information from sequence sets could be used. Perhaps by simplifying sequence space through the identification of key residues, and by simplifying conformation space as in the lattice methods, it will be possible to develop algorithms to generate a limited number of trial structures. These trial structures could then, in turn, be evaluated by further experiments and more sophisticated energy calculations.

REFERENCES AND NOTES

1. C. J. Epstein, R. F. Goldberger, C. B. Anfinsen, *Cold Spring Harbor Symp. Quant. Biol.* **28**, 439 (1963); C. B. Anfinsen, *Science* **181**, 223 (1973).
2. R. E. Dickerson, *Sci. Am.* **242**, 136 (March 1980).
3. M. D. Hampsey, G. Das, F. Sherman, *FEBS Lett.* **231**, 275 (1988).
4. D. Bashford, C. Chothia, A. M. Lesk, *J. Mol. Biol.* **196**, 199 (1987).
5. A. M. Lesk and C. Chothia, *ibid.* **136**, 225 (1980).
6. M. F. Perutz, J. C. Kendrew, H. C. Watson, *ibid.* **13**, 669 (1965).
7. C. Chothia and A. M. Lesk, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399 (1965).
8. J. U. Bowie and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152 (1989).
9. J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241**, 53 (1988); *Proteins Struct. Funct. Genet.*, in press.
10. D. Shortle, *J. Biol. Chem.* **264**, 5315 (1989).
11. J. H. Miller *et al.*, *J. Mol. Biol.* **131**, 191 (1979).
12. S. Sprang *et al.*, *Science* **237**, 905 (1987); C. S. Craik, S. Roczniak, C. Largman, W. J. Rutter, *ibid.*, p. 909.
13. H. C. M. Nelson and R. T. Sauer, *J. Mol. Biol.* **192**, 27 (1986).
14. M. H. Hecht, J. M. Sturtevant, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5685 (1984).
15. T. Alber, D. Sun, J. A. Nye, D. C. Muchmore, B. W. Matthews, *Biochemistry* **26**, 3754 (1987).
16. D. Shortle and A. K. Meeker, *Proteins Struct. Funct. Genet.* **1**, 81 (1986).
17. A. M. Lesk and C. Chothia, *J. Mol. Biol.* **160**, 325 (1982).
18. W. R. Taylor, *ibid.* **188**, 233 (1986).
19. W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959); R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8069 (1986).
20. W. A. Lim and R. T. Sauer, *Nature* **339**, 31 (1989); in preparation.
21. Lesk and Chothia (*5*) have argued that a protein core composed solely of hydrogen-bonded residues would also be inviable on evolutionary grounds, as a mutational change in one core residue would require compensating changes in any interacting residue or residues to maintain a stable structure.
22. T. M. Gray and B. W. Matthews, *J. Mol. Biol.* **175**, 75 (1984); E. N. Baker and R. E. Hubbard, *Prog. Biophys. Mol. Biol.* **44**, 97 (1984).
23. F. M. Richards, *J. Mol. Biol.* **82**, 1 (1974).
24. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987).
25. J. T. Kellis, Jr., K. Nyberg, A. R. Fersht, *Biochemistry* **28**, 4914 (1989); W. S. Sandberg and T. C. Terwilliger, *Science* **245**, 54 (1989).
26. A. A. Pakula and R. T. Sauer, *Proteins Struct. Funct. Genet.* **5**, 202 (1989).
27. B. C. Cunningham and J. A. Wells, *Science* **244**, 1081 (1989); R. M. Breyer and R. T. Sauer, *J. Biol. Chem.* **264**, 13348 (1989).
28. B. C. Cunningham, P. Jhurani, P. Ng, J. A. Wells, *Science* **243**, 1330 (1989).
29. L. H. Pearl and W. R. Taylor, *Nature* **329**, 351 (1987).
30. W. J. Brown *et al.*, *J. Mol. Biol.* **42**, 65 (1969); J. Greer, *ibid.* **153**, 1027 (1981); J. M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 99 (1988).
31. W. R. Taylor, *Protein Eng.* **2**, 77 (1988).
32. M. A. Navia *et al.*, *Nature* **337**, 615 (1989).
33. M. Schiffer and A. B. Edmundson, *Biophys. J.* **7**, 121 (1967); V. I. Lim, *J. Mol. Biol.* **88**, 857 (1974); *ibid.*, p. 873.
34. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Nature* **299**, 371 (1982); D. Eisenberg, D. Schwarz, M. Komaromy, R. Wall, *J. Mol. Biol.* **179**, 125 (1984); D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 140 (1984).
35. T. R. Burglin, *Cell* **53**, 339 (1988).
36. G. Otting *et al.*, *EMBO J.* **7**, 4305 (1988).
37. J. N. Breg, R. Boelens, A. V. E. George, R. Kaptein, *Biochemistry* **28**, 9826 (1989); M. G. Zagorski, J. U. Bowie, A. K. Vershon, R. T. Sauer, D. J. Patel, *ibid.*, p. 9813.
38. R. M. Sweet and D. Eisenberg, *J. Mol. Biol.* **171**, 479 (1983).
39. J. U. Bowie, N. D. Clarke, C. O. Pabo, R. T. Sauer, *Proteins Struct. Funct. Genet.*, in preparation.
40. K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
41. A. Sikorski and J. Skolnick, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2668 (1989); A. Kolinski, J. Skolnick, R. Yaris, *Biopolymers* **26**, 937 (1987); D. G. Covell and R. L. Jernigan, *Biochemistry*, in press.
42. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
43. S. R. Jordan and C. O. Pabo, *Science* **242**, 893 (1988).
44. R. M. Breyer, thesis, Massachusetts Institute of Technology, Cambridge (1988).
45. J.-L. Fauchere and V. Pliska, *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369 (1983).
46. We thank C. O. Pabo and S. Jordan for coordinates of the NH_2-terminal domain of λ repressor and its operator complex. We also thank P. Schimmel for the use of his graphics system and J. Burnbaum and C. Francklyn for assistance. Supported in part by NIH grant AI-15706 and predoctoral grants from NSF (J.R.-O.) and Howard Hughes Medical Institute (W.A.L.).

rticle No. jmbi.1998.2362 available online at http://www.idealibrary.com on

JMB

Effect of the Extra N-terminal Methionine Residue on the Stability and Folding of Recombinant α-Lactalbumin Expressed in *Escherichia coli*

Tapan K. Chaudhuri[1], Katsunori Horii[2], Takao Yoda[1], Munehito Arai[1] Shinji Nagata[3], Tomoki P. Terada[1], Hidefumi Uchiyama[4], Teikichi Ikura[1] Kouhei Tsumoto[2], Hiroshi Kataoka[3], Masaaki Matsushima[5] Kunihiro Kuwajima[1]* and Izumi Kumagai[2]

[1]Department of Physics
Graduate School of Science
University of Tokyo, Tokyo
113-0033, Japan

[2]Department of Biomolecular
Engineering, Tohoku
University, Sendai
980-8579, Japan

[3]Department of Biotechnology
Faculty of Agriculture & Life
Sciences, University of Tokyo
Tokyo, 113-8658, Japan

[4]Department of Biological
Science and Technology
Science University of Tokyo
2941 Yamazaki, Noda, Chiba
278-8510, Japan

[5]Rational Drug Design
Laboratories, Fukushima
960-1242, Japan

**Corresponding author*

The structure, stability, and unfolding-refolding kinetics of *Escherichia coli*-expressed recombinant goat α-lactalbumin were studied by circular dichroism spectroscopy, X-ray crystallography, and stopped-flow measurements, and the results were compared with those of the authentic protein prepared from goat milk. The electric properties of the two proteins were also studied by gel electrophoresis and ion-exchange chromatography. Although the overall structures of the authentic and recombinant proteins are the same, the extra methionine residue at the N terminus of the recombinant protein remarkably affects the native-state stability and the electric properties. The native state of the recombinant protein was 3.5 kcal/mol less stable than the authentic protein, and the recombinant protein was more negatively charged than the authentic one. The recombinant protein unfolded 5.7 times faster than the authentic one, although there were no significant differences in the refolding rates of the two proteins. The destabilization of the recombinant protein can be fully interpreted in terms of the increased unfolding rate of the protein, indicating that the N-terminal region remains unorganized in the transition state of refolding, and hence is not involved in the folding initiation site of the protein. A comparison of the X-ray structures of recombinant α-lactalbumin determined here with that of the authentic protein shows that the structural differences between the proteins are confined to the N-terminal region. Theoretical considerations for the differences in the conformational and solvation free energies between the proteins show that the destabilization of the recombinant protein is primarily due to excess conformational entropy of the N-terminal methionine residue in the unfolded state, and also due to less exposure of hydrophobic surface on unfolding. The results suggest that when the N-terminal region of a protein has a rigid structure, expression of the protein by *E. coli*, which adds the extra methionine residue, destabilizes the native state through a conformational entropy effect. It also shows that differences in the electrostatic interactions of the N-terminal amino group with the side-chain atoms of Thr38, Asp37, and Asp83 bring about a difference in the pK_a value of the N-terminal amino group between the proteins, resulting in a greater negative net charge of the recombinant protein at neutral pH.

Keywords: recombinant goat α-lactalbumin; extra N-terminal methionine residue; protein folding; X-ray crystallographic study; conformational entropy

Abbreviations used: GdnHCl, guanidine hydrochloride; CD, circular dichroism; N, native; U, unfolded; SDS, sodium dodecyl sulfate; CNBr, cyanogen bromide; UV, ultraviolet; ASA, accessible surface area.

E-mail address of the corresponding author: kuwajima@phys.s.u-tokyo.ac.jp

/
-,
).
t-
ol.
J.
:at-
res
3io-
lera
ıbile
from
78.
L. &
cell-
coli.
. K.,
1992).
and B
ıerichia
in vivo.
(1981).
ubunits
d those
41-346.
hic evi-
·N, and
Soc. 104,
s, J., De
rs, O. M.,
?. (1995).
ɔs of ana-
ı) adeno-
ıanosomal
dehydro-
., Connell,
ı6). Crystal
:in, LT-IIb.
H. R. &
sequence of
ɔsylated by
98.
:er, F., Kim,
, Sarfaty, S.,
ein crystallo-
n Sci. 3(10),
L., Nance, S.,
tbrook, E. M.
structure of
73.
by F. Cohen

Introduction

The N-terminal sequence of a recombinant protein expressed in *Escherichia coli* is known to start with formyl-methionine (Marcker & Sanger, 1964), which is in most cases subsequently processed by deformylase enzyme (Adams, 1968; Takeda & Webster, 1968), and removed by methionine aminopeptidase to finally produce the N-terminal methionine-free recombinant protein. However, removal of the N-terminal methionine does not always take place, and about half of *E. coli*-expressed proteins contain the extra N-terminal methionine residue, because the aminopeptidase action depends on the nature of the penultimate amino acid residue (Moerschell *et al.*, 1990). Therefore, the effect of the N-terminal methionine residue, when present, on the structure, stability and folding of *E. coli*-expressed recombinant proteins should be an important issue in biophysical and molecular biological studies that use such recombinant proteins, although this has not been taken seriously in most cases.

The biological and physicochemical properties of the methionylated proteins expressed in *E. coli* may differ from those of the authentic proteins that do not have the N-terminal methionine. For example, recombinant hen egg-white lysozyme contains the N-terminal methionine residue (Miki *et al.*, 1987; Mine *et al.*, 1997) and has lower solubility and stability than the authentic form (Imoto *et al.*, 1987). Similarly, recombinant apomyoglobin expressed in *E. coli* contains the extra N-terminal methionine residue and is less stable than the authentic protein (Hargrove *et al.*, 1994). On the other hand, the presence of the extra N-terminal methionine or the extension or truncation of the N-terminal residues does not interfere with the native-state stability in certain other globular proteins (Kordel *et al.*, 1989; Duverger *et al.*, 1991). In recombinant ribonuclease A, the extra N-terminal methionine is even known to stabilize the native structure (Schultz & Baldwin, 1992; Aronsson *et al.*, 1995). However, details of the effects of the extra N-terminal methionine residue on the structure, stability, and folding of the proteins have not yet been well understood.

α-Lactalbumin is a milk Ca^{2+}-binding protein, which consists of 123 amino acid residues and has a molecular mass of 14,200 Da. The three-dimensional structure of α-lactalbumin from several mammalian species, including goat, cow, guinea pig, and human, has been determined by X-ray crystallographic analysis (Acharya *et al.*, 1991; Pike *et al.*, 1996), and it is very similar to the structure of c-type lysozyme, a homologous protein. α-Lactalbumin has been used actively as a model protein in studies of protein folding (Sugai & Ikeguchi, 1994; Kuwajima, 1989, 1996; Vanderheeren & Hanssens, 1994; Uchiyama *et al.*, 1995; Schulman & Kim, 1996; Arai & Kuwajima, 1996; Schulman *et al.*, 1997; Wilson *et al.*, 1996; Shimizu *et al.*, 1996; Balbach *et al.*, 1996; Katsumata *et al.*, 1996; Kataoka *et al.*, 1997; Kuhlman *et al.*, 1997; Wu & Kim, 1997; Pfeil, 1998; Ikeguchi *et al.*, 1998), because this protein readily adopts a molten globule state, which is known to be identical with a folding intermediate (Kuwajima, 1989, 1996; Ptitsyn, 1995), under a variety of conditions, including those at a low pH, at a moderate concentration of guanidine hydrochloride (GdnHCl), and in the absence of Ca^{2+} and other salts (Kuwajima, 1989, 1996). Recombinant α-lactalbumin expressed in *E. coli*, though containing the extra N-terminal methionine, has often been used in these studies of protein folding. A recent study has, however, shown that like recombinant hen egg-white lysozyme, recombinant bovine α-lactalbumin is less stable than the authentic protein, although the lactose synthase regulatory activities of the recombinant and authentic proteins have been shown to be identical with each other (Ishikawa *et al.*, 1998).

Here, we show that *E. coli*-expressed recombinant goat α-lactalbumin is destabilized by the presence of the extra N-terminal methionine residue by as much as 3.5 kcal/mol and has a more negative electric net charge than the authentic protein. It is concluded that the destabilization of the recombinant protein is primarily brought about by an extra conformational entropy of the methionyl residue in the unfolded state and that the more negative charge of the recombinant protein is caused by a decrease in the pK_a value of the N-terminal amino group. Because the N-terminal methionine remarkably destabilizes recombinant α-lactalbumin, the role of the N terminus in the folding of this protein has also been investigated by stopped-flow circular dichroism (CD) studies of the unfolding and refolding kinetics of the recombinant and authentic proteins. The destabilization of the recombinant protein is shown to be entirely interpreted in terms of an increase in the unfolding rate, indicating that the N terminus is not involved in the folding initiation site of α-lactalbumin.

Results

Structure of folded recombinant goat α-lactalbumin

The recombinant wild-type protein was expressed in *E. coli* as inclusion bodies with a high yield (15 mg per litre of culture). The protein was solubilized in 8 M urea and refolded in a redox buffer in the absence of urea at pH 8.5 and 4 °C. The process of refolding was monitored by reversed-phase HPLC (Uchiyama *et al.*, 1995), and the folded protein was purified (see Materials and Methods). The peptide and aromatic CD spectra of the recombinant protein were measured under native conditions (0 M GdnHCl (pH 8.0) at 25 °C), and compared with those of authentic goat α-lactalbumin (Figure 1(a) and (b)). There is no significant difference in the CD spectra between the proteins in the aromatic and peptide regions, so that the secondary and tertiary structures of the

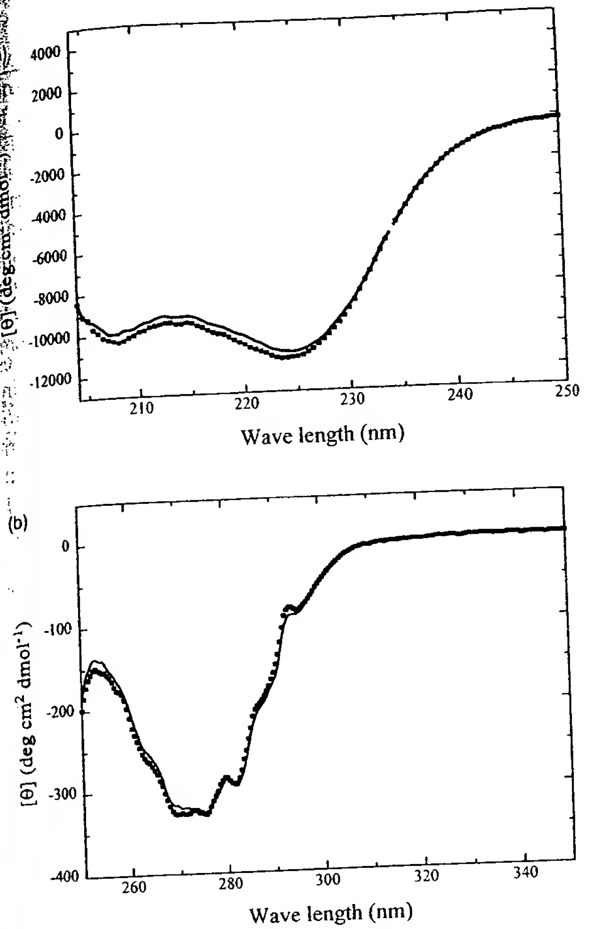


Figure 1. (a) Far and (b) near-UV CD spectra of authentic and recombinant goat α-lactalbumin measured in the presence of 1 mM $CaCl_2$ at pH 7.0 and 25 °C. The continuous line denotes the authentic protein and the filled squares denote the recombinant protein.

two proteins are essentially identical with each other. This conclusion is confirmed by the X-ray crystallographic structure of recombinant goat α-lactalbumin (see below). The results thus indicate that the folded recombinant protein is correctly folded into the native structure. A study has also shown that the lactose synthase regulatory activity of the folded recombinant protein is the same as that of authentic α-lactalbumin (Uchiyama *et al.*, 1995).

Equilibrium unfolding

The GdnHCl-induced equilibrium unfolding transition of the folded recombinant protein was studied by the peptide and aromatic CD spectra, and the results were compared with those of authentic goat α-lactalbumin. Figure 2 shows the unfolding transition curves of the two proteins measured by the CD ellipticities, at 222 and 270 nm, and these ellipticities are expressed by the apparent fractional extent (F_{app}) of unfolding as a

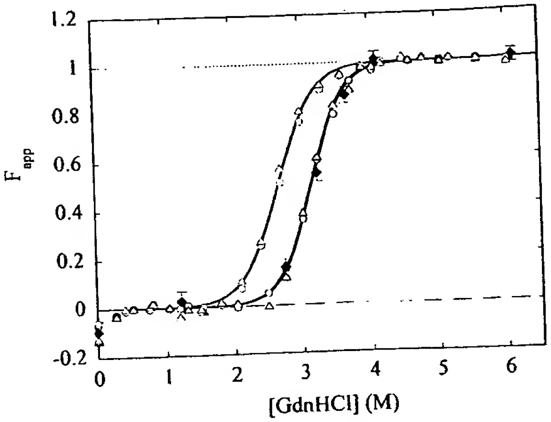


Figure 2. GdnHCl-induced unfolding transition curves for authentic, recombinant, and methionine-free goat α-lactalbumin. The unfolding was carried out at 25 °C in the presence of 1 mM $CaCl_2$, 50 mM NaCl, and 50 mM sodium cacodylate (pH 7.0), and the transitions were monitored by both far and near-UV CD measurements. Apparent fractions of unfolded species (F_{app}) were plotted against the concentration of GdnHCl. Open black circles and open red cirles denote the F_{app} values measured at 222 nm for authentic and recombinant proteins, respectively; open black triangles and open red triangles represent the F_{app} values measured at 270 nm for authentic and recombinant proteins, respectively; and the F_{app} values of the methionine-free recombinant protein measured at 222 nm are presented by filled blue diamonds.

From Figure 2, the unfolding transition curves measured at 222 and 270 nm are coincident with each other in authentic and recombinant α-lactalbumin, indicating that the unfolding transitions of the two proteins are well represented by a two-state mechanism, in which only the native (N) and the fully unfolded (U) states are populated in the transition zone as:

$$N \overset{K_U}{\rightleftharpoons} U \qquad (1)$$

Here K_U is the equilibrium constant of unfolding and relates to the free energy change, ΔG_U, of the unfolding transition as:

$$K_U = \exp(-\Delta G_U/RT) \qquad (2)$$

where R and T are the gas constant and the absolute temperature, respectively, and ΔG_U is assumed to be linearly dependent on GdnHCl concentration (C) as:

$$\Delta G_U = \Delta G_U^{H_2O} - mC = m(C_m - C) \qquad (3)$$

where $\Delta G_U^{H_2O}$ is the ΔG_U in the absence of the denaturant, C_m is the C at the midpoint of the unfolding transition, and m represents the dependence of ΔG_U on C and is a measure of the cooperativity of the transition (Pace, 1986). From

Table 1. Equilibrium unfolding transition parameters of goat α-lactalbumin

Name of protein	$\Delta G_U^{H_2O}$ (kcal/mol)	m (kcal/mol M)	C_m (M)	$\Delta\Delta G_U^{H_2O}$ (kcal/mol)	$\Delta\Delta G_U$ (kcal/mol) at 3.2 M GdnHCl
Authentic goat α-lactalbumin	13.8 ± 0.7	4.4 ± 0.2	3.15 ± 0.01	-	-
Recombinant goat α-lactalbumin	10.4 ± 0.5	3.9 ± 0.2	2.67 ± 0.01	−3.5	−1.9

by F_{app} is given as a function of C as:

$$F_{app}(C) = \frac{\exp[-m(C_m - C)/RT]}{1 + \exp[-m(C_m - C)/RT]} \qquad (4)$$

The values of m, C_m, and hence $\Delta G_U^{H_2O}$, for recombinant and authentic α-lactalbumin were calculated from the data of Figure 2 by the non-linear least-squares method. The unfolding parameters m, C_m, and hence $\Delta G_U^{H_2O}$, thus obtained are summarized in Table 1. The continuous lines in Figure 2 are the curves theoretically drawn with the parameter values of Table 1, and show excellent agreement between theory and the experimental data.

Figure 2 also shows that the unfolding transition of the recombinant protein occurs at a remarkably lower concentration of GdnHCl ($C_m = 2.7$ M) than the transition of authentic α-lactalbumin ($C_m = 3.2$ M). The difference in ΔG_U ($\Delta\Delta G_U$) is −3.5 kcal/mol at 0 M GdnHCl and −1.9 kcal/mol at 3.2 M GdnHCl, which is the C_m for the authentic protein (Table 1). Therefore, the folded recombinant protein is remarkably less stable than authentic α-lactalbumin, although their native structures are practically identical as evidenced by the CD spectra and X-ray structural analysis.

Gel electrophoresis and ion-exchange chromatography

In order to investigate further differences between recombinant and authentic α-lactalbumin, the electrophoretic and ion-exchange chromatographic behavior of the two proteins were investigated. Figure 3(a) shows electrophoretic patterns in a non- denaturing polyacrylamide gel at pH 9.4. It can be seen that the electrophoretic mobility of the recombinant protein is significantly greater than that of the authentic protein. Figure 3(b) shows the elution profiles of recombinant and authentic α-lactalbumin in an anion-exchange HPLC using a RESOURCE™ Q column (Pharmacia Biotech) with a linear gradient from 0 M to 0.5 M NaCl in the presence of 10 mM NaH_2PO_4-Na_2HPO_4 buffer (pH 7.0). The retention time is longer for the recombinant protein (22.9 minutes) than for the authentic one (19.6 minutes). Both of these results indicate that the recombinant protein is more negatively charged. These differences in the electric properties of the two proteins, however, disappear in the U state in 8 M urea. The electrophoretic mobilities and the chromatographic retention times of the proteins were found to be identical in the presence of 8 M urea (data not shown). Therefore,

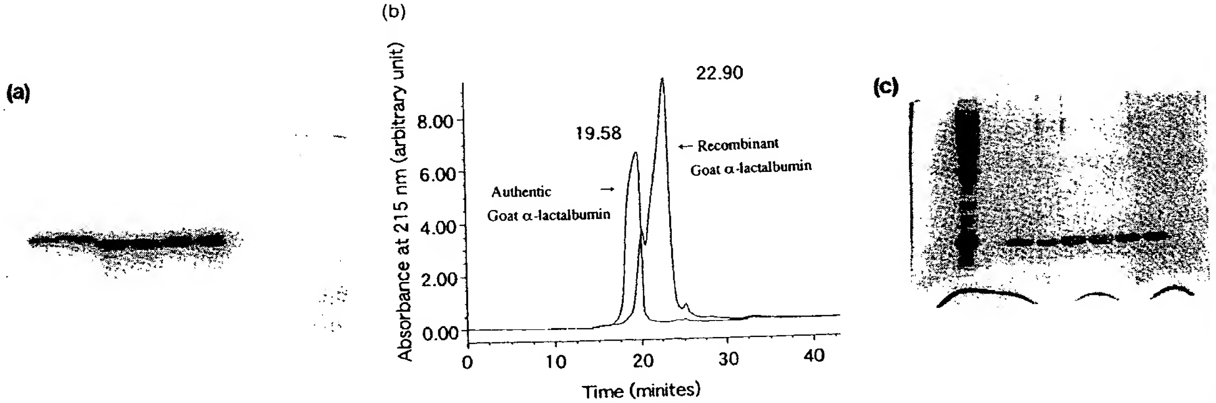


Figure 3. (a) Non-denaturing polyacrylamide gel electrophoresis of authentic and recombinant goat α-lactalbumin. The electrophoresis was carried out at pH 8.9 using 12% (w/v) acrylamide gel at room temperature. About 5 µg protein samples were applied in each lane. Lanes 1 and 2 contain authentic protein, lanes 3 and 4 contain recombinant protein, and lanes 5 and 6 contain an equimolecular mixture of authentic and recombinant protein. Lanes are numbered from left to right in (a) and (c). (b) Superimposed HPLC pattern of authentic and recombinant goat α-lactalbumin. The HPLC was performed with a Resource-Q anion exchange column at pH 7.0 using a linear gradient of 0 M-0.5 M NaCl containing 10 mM NaH_2PO_4-Na_2HPO_4 buffer. A 50 µl protein sample containing 50 and 60 µg of native and recombinant protein, respectively, was applied in the HPLC column. (c) SDS/polyacrylamide gel electrophoresis of authentic and recombinant goat α-lactalbumin. The electrophoresis was carried out using 15% (w/v) acrylamide gel at room temperature. Approximately 10 µg of protein was applied in each lane. Lane 1 contains low molecular mass marker proteins, lane 2 is blank, lanes 3 and 4 contain the authentic protein, lanes 5 and 6 contain the recombinant protein and lanes 7 and 8 contain an equimolecular mixture of the authentic and recombinant proteins.

the difference in the electric charge between the proteins must be caused by the structural folding of the proteins into the native structure.

SDS/polyacrylamide gel electrophoresis was also carried out for the recombinant and authentic proteins using 15% acrylamide in the resolving gel (Figure 3(c)). The electrophoretic mobilities of the two proteins are the same within the experimental error, indicating that there is no significant difference in the molecular mass between the proteins.

N-terminal sequence and mass spectrometric analyses

In order to identify any differences in the amino acid sequence, we performed N-terminal sequencing and mass spectrometric analysis of the recombinant and authentic proteins. The N-terminal sequences of the first five residues of the two proteins have shown that recombinant α-lactalbumin contains an additional methionine residue. The results of the mass spectrometric analysis indicate that the difference in mass between the recombinant and authentic proteins is 133 (Figure 4), which is nearly equal to the mass of a single methionine residue (131.19), confirming the presence of the extra methionine residue in the recombinant protein. Therefore, the only chemical difference that brings about the difference in the electric charge between the two proteins in the N state is the presence or absence of the extra methionine residue at the N terminus, and this difference may also lead to the remarkable difference in stability between the proteins.

Methionine-free recombinant α-lactalbumin

In order to directly investigate the effect of the extra methionine residue on the electric properties and stability of the recombinant protein, methionine-free recombinant α-lactalbumin was prepared by cyanogen bromide (CNBr) cleavage. Because there is no methionine residue in authentic goat α-lactalbumin, only the extra N-terminal methionine of the recombinant protein is expected to be removed by the CNBr cleavage. The removal of the methionine was confirmed by N-terminal sequencing and mass spectrometric analysis (data not shown). The absence of other cleavage products was confirmed by SDS/polyacrylamide gel electrophoresis. The near and far-UV CD spectra of the methionine-free recombinant protein overlap with those of the authentic and original recombinant proteins (data not shown). The electrophoretic mobility in the native gel and the retention time for the anion-exchange chromatography were found to be identical with those of the authentic protein (data not shown). The stability of the methionine-free recombinant protein against the GdnHCl-induced unfolding was investigated, and the equilibrium unfolding transition of the methionine-free protein is shown in Figure 2. The unfolding transition curve coincides well with that of the authentic protein, and gives the same C_m and $\Delta G_U^{H_2O}$ values. As a control, the authentic protein was also subjected to the conditions of CNBr cleavage, and it was confirmed that the unfolding transition of the protein was not affected by the cleavage conditions (data not shown). These results thus clearly indicate that the observed destabilization and the difference in the electric charge of the recombinant protein is solely due to the presence of the extra N-terminal methionine residue.

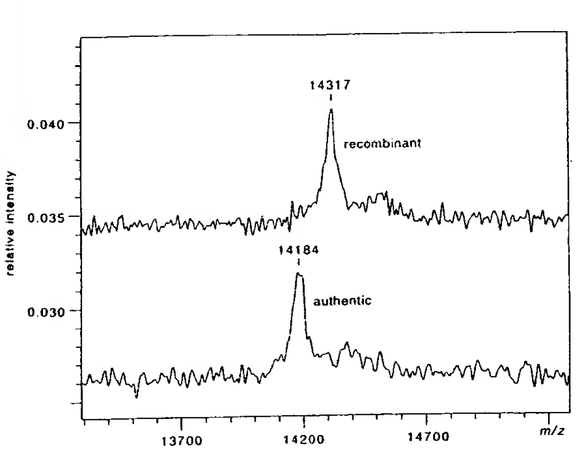


Figure 4. MALDI-TOF-MS mass spectroscopic pattern of authentic and recombinant goat α-lactalbumin. The upper trace is for recombinant and the lower one is for authentic protein.

Kinetics of refolding and unfolding

The above results indicate that the presence of the extra methionine residue at the N terminus of the recombinant protein decreases the relative stability of the N state by as much as 3.5 kcal/mol. Thus, it appears that both recombinant and authentic goat α-lactalbumin are useful for investigating the role of the N-terminal residue in the kinetic folding of α-lactalbumin. The kinetic unfolding and refolding reactions of the recombinant and authentic proteins were investigated by stopped-flow CD measurements. The unfolding and refolding reactions were induced by concentration jumps of GdnHCl from 1.0 to 5.4 M and from 5.5 to 0.5 M, respectively. The reactions were monitored by the ellipticity change at 225 nm at pH 7.0 and 25 °C. The kinetic progress curves for unfolding and refolding are shown in Figure 5(a) and (b), respectively, and the data were fitted by the non-linear least-squares method with the equation:

$$A(t) = A(\infty) + \Delta A_{\text{obs}} \sum \alpha_i \exp(-k_i t) \qquad (5)$$

where $A(t)$ and $A(\infty)$ are the observed values of the ellipticity at time t and infinite time, respectively, ΔA_{obs} is the observed total amplitude $[A(0) - A(\infty)]$, and k_i and α_i are the apparent first-

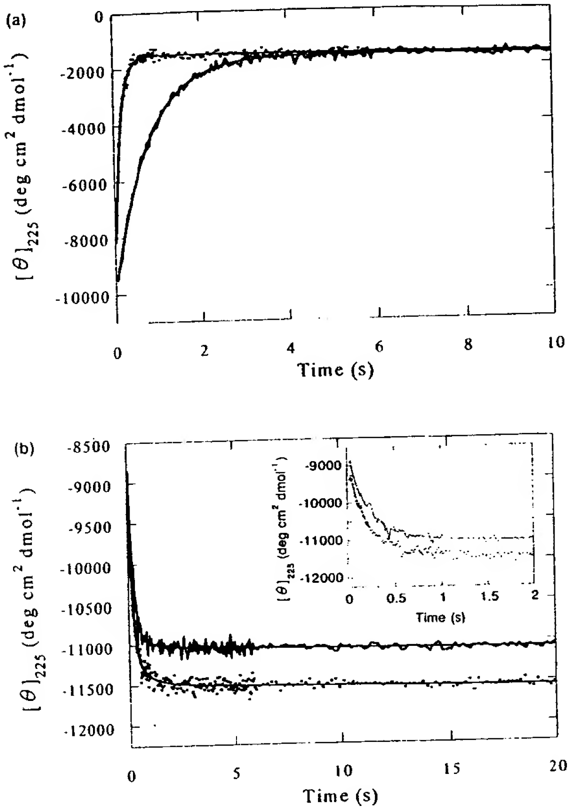


Figure 5. GdnHCl-induced (a) unfolding and (b) refolding kinetic progress curves of authentic and recombinant goat α-lactalbumin. The unfolding was initiated by a concentration jump from 1.0 M to 5.4 M and the refolding process was initiated by a concentration jump of 5.5 M to 0.5 M at 25°C in the presence of 1 mM $CaCl_2$, 50 mM NaCl, and 50 mM sodium cacodylate, pH 7.0, and the refolding and unfolding kinetics were monitored by the measurement of CD ellipticity at 225 nm using stopped-flow CD. The continuous line denotes authentic protein and the filled squares denote recombinant protein. (b) The inset shows the refolding progress curve within two seconds and the same notations are used for the transition curves. Theoretical kinetic progress curves are also shown in (a) and (b).

Table 2. Kinetic unfolding parameters of goat α-lactalbumin

Name of protein	k_1 (s^{-1})	$\alpha_1\Delta A_{obs}$ (deg cm^2 dmol^{-1})
Authentic goat α-lactalbumin	1.26 ± 0.01	−8384
Recombinant goat α-lactalbumin	7.18 ± 0.08	−8056

order rate constant and fractional amplitude, respectively, of the *i*th kinetic phase.

The kinetic progress curves for unfolding for both the recombinant and authentic proteins were well fitted to a single-exponential equation, and the apparent rate constants and the amplitudes for the two proteins are presented in Table 2. The kinetic progress curves for refolding were well fitted to the three-exponential equation, and the rate constants and the amplitudes are presented in Table 3. The unfolding reaction of recombinant α-lactalbumin is 5.7-times faster than that of the authentic protein, while there are no significant differences in the rate constants for the triphasic refolding reactions of the two proteins. Thus, it appears that the N-terminal end of goat α-lactalbumin is not essential for the kinetic folding of this protein (see Discussion).

X-ray crystallographic study

In order to further investigate the differences in the folded structure between recombinant and authentic goat α-lactalbumin, an X-ray crystallographic analysis of the recombinant protein was performed, and the structure was compared with that reported for the authentic protein structure. The crystallographic data are summarized in Table 4. The space group of the crystal of the recombinant protein was altered to $P2_12_12$ from $P2_1$ in which the authentic protein was packed (Pike *et al*., 1996). The number of protein molecules in the asymmetric unit was one, although there were two (Mol A and Mol B) in the authentic protein crystal. The final *R* and free *R* factors were 0.191 and 0.278 in the resolution range of 8.0 to 2.0 Å. The overall error was estimated at 0.19 Å by a Luzzati plot (Luzzati, 1952). As the space group is altered in the recombinant protein crystal, the N-terminal methionine may affect the molecular packing in the crystal. However, the interactions between the two independent authentic molecules (Mol A and Mol B) were found to be very similar to the interactions between the symmetry-related recombinant molecules (Figure 6).

The structural differences between the recombinant and the authentic proteins are shown in Figure 7, which represents the distances between the C^α atoms of the two molecules. The root-mean-square deviations of the main-chain atoms are 0.55 Å between the recombinant protein molecule

Table 3. Kinetic refolding parameters of goat α-lactalbumin

Name of protein	k_1 (s^{-1})	$\alpha_1\Delta A_{obs}$ (deg cm^2 dmol^{-1})	k_2 (s^{-1})	$\alpha_2\Delta A_{obs}$ (deg cm^2 dmol^{-1})	k_3 (s^{-1})	$\alpha_3\Delta A_{obs}$ (deg cm^2 dmol^{-1})
Authentic goat α-lactalbumin	0.11 ± 0.05	64.28	1.3 ± 1.1	145	4.9 ± 0.3	2282
Recombinant goat α-lactalbumin	0.09 ± 0.04	70.8	1.3 ± 0.4	335	5.7 ± 0.4	2234

Table 4. Crystallization, data collection, and refinement statistics of recombinant goat α-lactalbumin

A. Crystallization	
Reservoir solution	1.0 mM $CaCl_2$ 16-20% PEG8000 0.05 M KH_2PO_4 pH 6.0
Protein concentration	20 mg protein/ml
Temperature (°C)	20
B. Crystal data	
Space group	$P2_12_12$
a, *b*, *c* (Å)	44.9, 88.9, 32.2
In an asymmetric unit	1
X-ray generator	Cu target (4.5 kW)
Resolution at measurements (Å)	1.75
Total number of ind. refl.	12,533
R_{merge}[a]	0.069
Completeness (%)	92.5
C. Structure determination	
Method	Mol. replacement
Model structure	Baboon α-LA
Software	X-PLOR 3.1
D. Refinement	
Software	X-PLOR 3.1
Resolution range (Å)	8.0-2.0
R-factor[b]	0.191
R_{free}[c]	0.278
Rms deviations in:	
Bond length (Å)	0.010
Bond angles (°)	1.554

[a] $R_{merge} = \Sigma_h \Sigma_i |I(h, i) - \langle I(h)\rangle| / \Sigma_h \Sigma_i I(h, i)$, where $I(h, i)$ is the intensity value of the *i*th measurement of *h* and $\langle I(h)\rangle$ is the corresponding mean value of $I(h)$ for all *i* measurements.

[b] $R\text{-factor} = \Sigma||F_{obs}| - |F_{calc}||/|F_{obs}|$, where $|F_{obs}|$ and $|F_{calc}|$ are observed and calculated structure factor amplitude respectively.

[c] R_{free} is the same as *R*-factor, but for a 10% subset of all reflections.

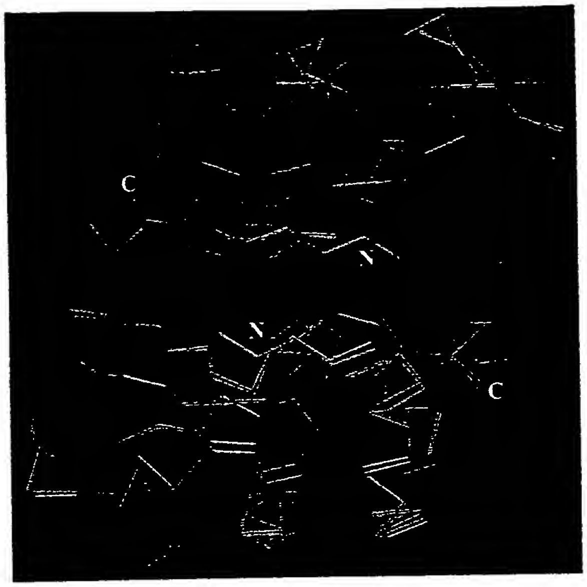


Figure 6. The molecular packings of recombinant and authentic goat α-lactalbumin in the crystals. The main-chain atoms of Mol A (orange) were superimposed on those of the recombinant protein molecule (blue). The same transformation matrixes were applied on Mol B (yellow). The Figure shows two of the symmetry-related recombinant protein molecules, and Mol A and B are overlaid. The space group was $P2_12_12$ in the recombinant protein crystal and $P2_1$ in the authentic protein crystal. The interactions in both crystals were very similar.

and Mol A, and 0.63 Å between the recombinant molecule and Mol B. These values are larger than the root-mean-square deviation between Mol A and Mol B (0.27 Å). From Figure 7, we can see that the intermolecular interactions remarkably affect the structure of the N-terminal and loop regions of the protein, especially between residues 105 and 110, but that the overall structures of the recombinant and authentic proteins are essentially identical, supporting previous observations of the same CD spectra of the proteins in solution.

The structures around the N termini of the recombinant protein and the two molecules of the authentic protein are shown in Figure 8, and we may see structural differences that give rise to the differences in the electric properties and stability between them. The N-terminal amino group strongly interacts with the side-chain atoms of Thr38 and Gln39 in Mol A (Figure 8(a)) and Thr38 in Mol B through hydrogen bonds and/or salt bridges (Figure 8(b)). A similar interaction can also be observed in the recombinant protein, in which the N-terminal amino group is bound to the side-chain of Gln39 by a hydrogen bond (Figure 8(c)), but this interaction may be significantly stronger than the corresponding interaction in the authentic protein (see Discussion). It can also be seen from Figure 8(c) that the methionine side-chain of the recombinant protein is directly in contact with the side-chain of Gln2, and that the orientation of the methionine side-chain is fixed by the hydrogen bonds between the N-terminal amino group and the side-chain of Gln39, and between the main-

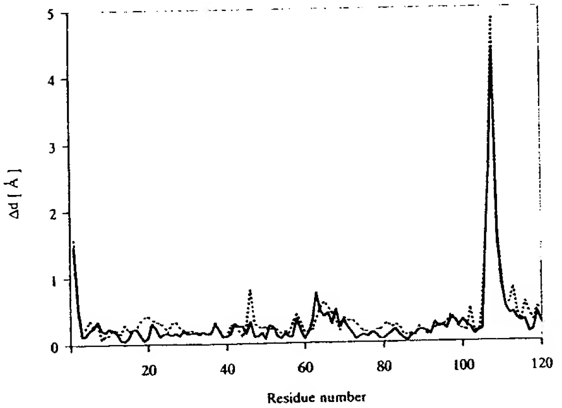


Figure 7. The structural differences between the corresponding C^α atoms of the recombinant and authentic protein molecules. Differences were observed in the N-terminal residues and the flexible loop residues of 105-110. The loop residues of the recombinant protein were affected by the neighboring molecules in the crystal.

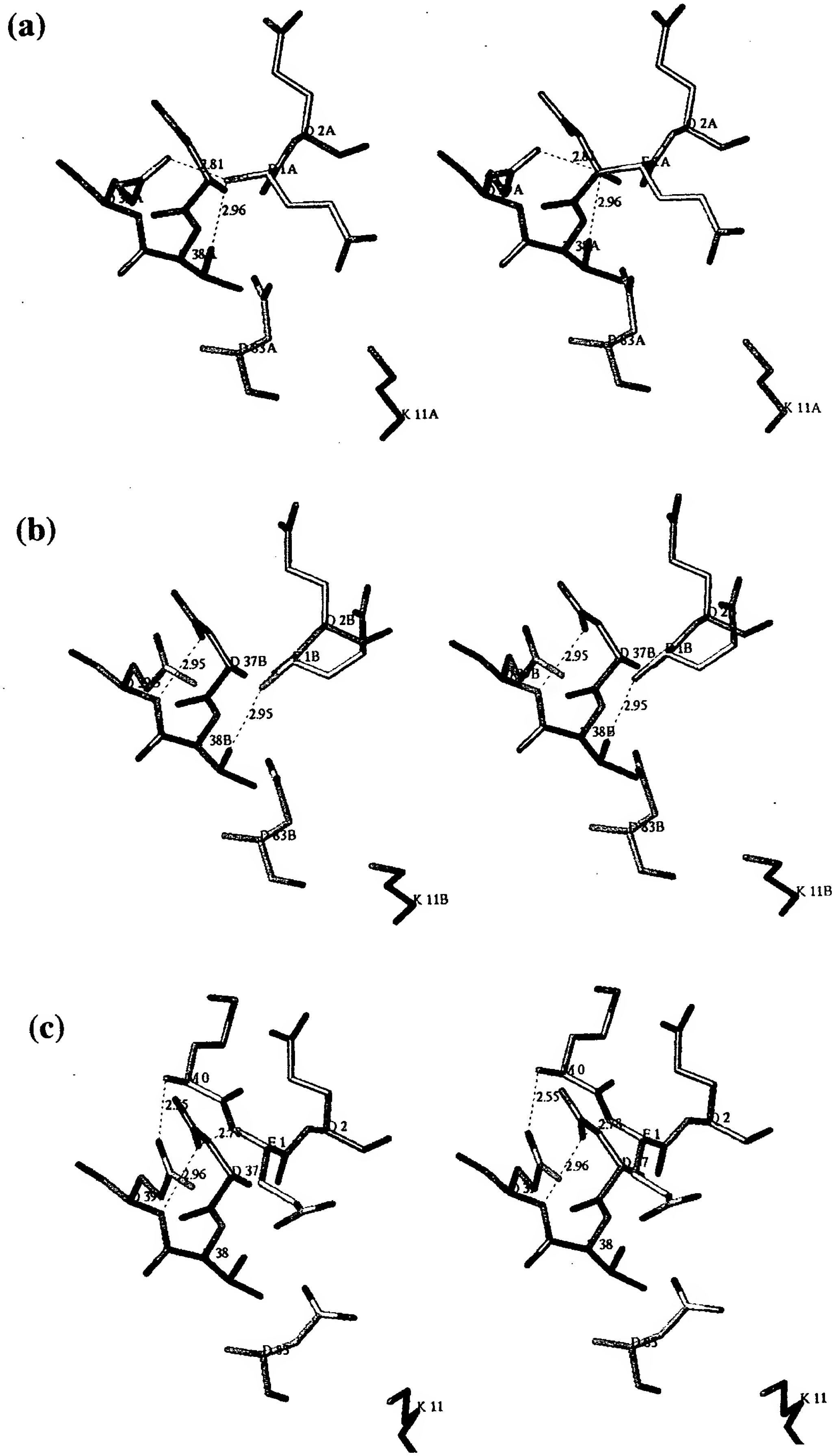


Figure 8 (*legend opposite*)

chain amido group of Glu1 and the carboxyl group of Asp37. The side-chains of the methionine and Gln2 residues weakly interact with a neighboring protein molecule by van der Waals contacts (Figure 8(c)). The main-chain conformations of residues Glu1 and Gln2 of the recombinant protein are almost the same as those of the authentic molecules. It is interesting that very similar intermolecular interactions are found in the recombinant and authentic protein molecules, although their crystallographic packings are different. The side-chain of Glu1 is folded into the inside of the recombinant molecule and interacts with the amino group of Lys11, and this conformation is similar to that in Mol A, but the Glu1 side-chain is shifted further from the amino group of Lys11 in the recombinant protein. The corresponding conformation of the Glu1 side-chain of Mol B is affected by the positively charged His107 side-chains of the symmetry-related molecules (Mol A and Mol B) in the crystal of the authentic protein. Without the presence of these positive charges near Glu1 of Mol B, the conformation of the side-chain is similar to that of the recombinant molecule and Mol A of the authentic protein.

Discussion

The present results show that the presence of the additional N-terminal methionine residue in recombinant α-lactalbumin expressed in *E. coli* remarkably decreases the stability of the native protein and increases the apparent net negative charge in the native state. Many proteins expressed in *E. coli* have the N-terminal methionine residue, although whether or not the methionine residue is present depends on the next residue in the recombinant protein (Miller *et al*., 1987). Thus, the effect of the N-terminal methionine on the structure, stability, and other properties of an *E. coli*-expressed protein is important when we use the recombinant protein in biophysical and molecular biological studies. Such effects of the N-terminal methionine residue have, however, been ignored so far in most cases. As far as we are aware, the present study is the first of its kind in which the effect of the N-terminal methionine residue is thoroughly investigated by equilibrium unfolding and kinetic unfolding-refolding studies as well as by the CD and X-ray crystallographic analyses. Only recently, Ishikawa *et al*. (1998) have reported the effect of the N-terminal methionine on the thermal unfolding of bovine α-lactalbumin, and their results and the present results should be complementary to each other.

Because the N-terminal methionine affects the native state stability of α-lactalbumin, the effects of the methionine on the kinetics of refolding and unfolding of the protein were investigated. The results of our study show that the recombinant protein unfolds 5.7 times faster than the authentic one, whereas the rates of refolding remain the same. The results should provide an insight into the role of the N terminus in the folding mechanism for α-lactalbumin.

We will first consider here the structural aspects that explain differences in the native state stability and the electric properties between the recombinant and authentic proteins on the basis of our CD and X-ray crystallographic data. We will then discuss the mechanism of folding for goat α-lactalbumin on the basis of the kinetic unfolding and refolding data.

Stabilities of recombinant and authentic α-lactalbumin

Structure around the N terminus

The results show that although recombinant and authentic α-lactalbumin follow the two-state unfolding transition (Figure 2), the recombinant protein is less stable than the authentic one by as much as 3.5 kcal/mol in the absence of GdnHCl. In order to understand this stability difference ($\Delta\Delta G_U$), we determined the X-ray structure of the recombinant protein, and this structure was compared with the X-ray structure of the authentic protein determined by Pike *et al*. (1996). The overall structures of the two proteins are essentially identical with each other, being consistent with the identical CD spectra of the proteins, and the structural differences between the proteins have been found to be localized in the N-terminal and the 105-110 loop regions (Figure 7). Because the structural differences in the 105-110 loop region, which is very flexible in the N state, are likely to be caused by a difference in the crystallographic packing between the proteins (Acharya *et al*., 1991; Harata & Muraki, 1992; Pike *et al*., 1996), we have concentrated our attention on the structural differences in the N-terminal region and investigated any interactions that are present in the authentic protein but missing in the recombinant one. Our data, however, show that there are no such interactions identified in the X-ray structures. From Figure 8, it can be seen that the N-terminal amino group of

Figure 8. Stereo views of the N-terminal region of authentic goat α-lactalbumin, (a) Mol A and (b) Mol B, and (c) recombinant goat α-lactalbumin. The main-chain structures of these three molecules were very similar. But the mainchain of the recombinant protein was shifted to the outside of the molecule, compared with those of the authentic molecules. The side-chain conformation of Gln2 of Mol B is different from those of the others. The side-chain of Asp83 of the recombinant protein was not clearly seen in the electron density map, and the *B*-factors of the side-chain atoms were high. Therefore, the model coordinates could not be explicitly determined. Certain distances are shown in Å, and the residues are shown by the one-letter code.

Glu1 of the authentic protein is hydrogen-bonded with two side-chain oxygen atoms of Thr38 and Gln39 in Mol A and with a side-chain oxygen atom of Thr38 in Mol B. A similar hydrogen bond is also observed in the recombinant protein between the N-terminal amino group and Gln39, and the length of the hydrogen bond is smaller than that in the authentic protein, suggesting that the hydrogen bond is even stronger in the recombinant protein (Figure 8(c)). Although the hydrogen bond between the N-terminal amino group and Thr38 is missing in the recombinant protein, there is an alternative hydrogen bond between the main-chain amido group of Glu1 and the carboxyl group of Asp37. The degrees of the packing interactions of the side-chain atoms are also very similar in the N-terminal regions of the two proteins. The side-chains are closely packed in both proteins. Furthermore, contributions of electrostatic interactions around the N termini to the destabilization of the recombinant protein will be shown to be negligibly small, although they are related to the difference in the electric net charge between the proteins (see below). Therefore, the observed destabilization cannot be interpreted in terms of the presumed interactions missing in the native structure of the recombinant protein.

Conformational entropy of the methionine residue and solvation free energies

If the destabilization of the recombinant protein cannot be simply explained by the interactions identified in the X-ray structures of the recombinant and authentic proteins, what makes the recombinant protein less stable? At this point, it should be noted that the N-terminal residues of both the recombinant and authentic proteins are involved in a rigid structure, so that all the atoms of the residues can be traced in the electron density maps of the proteins by X-ray crystallographic analysis. The *B*-factors of the backbone atoms of the N-terminal methionine residue of the recombinant protein were found to range from 31 to 35 Å^2. The values are much larger than those of the residues buried inside the protein molecule (8-15 Å^2), but are smaller than those of the fully exposed residues in flexible loop regions. This means that the presence of the additional methionine residue in the recombinant protein destabilizes the native state through an entropic effect, which arises from an additional conformational entropy of the methionine residue in the U state. Because the structure around the N terminus is rigid in the N state of the recombinant protein, the additional methionine residue leads to an increase in entropy on unfolding. Thus, the free energy change of unfolding (ΔG_U), which is the difference in the free energy between the N and U states, decreases, and hence the N state of the recombinant protein is destabilized. estimated at 20 cal/(mol K) by Oobatake & Ooi (1993) from an analysis of hydration and heat stability effects on the unfolding of 14 globular proteins, and this corresponds with the free energy change of −5.9 kcal/mol at 25 °C. This value is close to but lower than the observed difference ($\Delta\Delta G_U = -3.5$ kcal/mol) in ΔG_U between the recombinant and authentic proteins. We have, however, ignored the contribution of the hydration free energy, ΔG_h^u, and the enthalpic contribution of the conformational unfolding, ΔH_c^u, which mostly arises from the van der Waals interaction energy, to the $\Delta\Delta G_U$ (Oobatake & Ooi, 1993). These contributions are expected to be proportional to the change in the accessible surface area of the methionine residue on unfolding (Oobatake & Ooi, 1993) and may explain the above difference between the expected contribution of the conformational entropy ($-T\Delta S_c^u$) and the observed $\Delta\Delta G_U$. The values of ΔG_h^u and ΔH_c^u of the N-terminal methionine residue were calculated by the method described by Oobatake & Ooi (1993), and they were −1.2 and 3.3 kcal/mol for ΔG_h^u, and ΔH_c^u, respectively, so that the free energy change of unfolding of the methionine residue (ΔG^u) was estimated at −3.8 kcal/mol (see equation (8)), which was in good agreement with the observed $\Delta\Delta G_U$ (see Materials and Methods). The contribution of other residues to the $\Delta\Delta G_U$ was also estimated, and it was less than 1 kcal/mol (see Materials and Methods), confirming that the increase in the conformational entropy of the N-terminal methionine residue on unfolding is a dominant factor determining the $\Delta\Delta G_U$.

In the above argument of $\Delta\Delta G_U$, however, we have implicitly assumed that the U state is fully unfolded in both the recombinant and authentic proteins. Thus, if there is a difference in the U-state structure between the proteins, such a difference may also contribute to the $\Delta\Delta G_U$. In fact, the *m* value of the equilibrium unfolding transition is found to be smaller for the recombinant protein (Table 1). Lower values of *m* are usually thought to be due to less exposure of hydrophobic surface on unfolding. Because the native structure is essentially identical between recombinant and authentic α-lactalbumin, the less exposure of hydrophobic surface must be due to a difference in the U-state structure, and the U state of the recombinant protein less exposes the hydrophobic surface than that of the authentic one. Similar effects of hydrophobic replacements of amino acid residues on the U-state structure have also been reported in staphylococcal nuclease (Shortle, 1996). The less difference in solvent exposed hydrophobic surface means a smaller difference in ΔG_U. Therefore, this may also be a factor determining the $\Delta\Delta G_U$ between the recombinant and authentic proteins.

Comparison with other proteins

authentic bovine α-lactalbumin using thermal denaturation measurements of the proteins. They have shown that the destabilization of the recombinant protein is caused by an entropic effect because the enthalpy change of the thermal unfolding is the same for the two proteins, and their result is fully consistent with our proposal regarding the destabilization of the recombinant protein described above. Although Ishikawa *et al.* (1998) have attributed the destabilization of the recombinant protein to a weakening of the apparent Ca^{2+}-binding strength, this interpretation seems to be nothing more than a rephrasing of the destabilization of the protein because the apparent Ca^{2+}-binding strength of α-lactalbumin is known to be linked to the $N \rightleftharpoons U$ equilibrium of the apo protein (Hiraoka & Sugai, 1985). Our X-ray structural data show that there is no essential difference in the structure of the Ca^{2+}-binding site between the authentic and recombinant proteins, indicating that the weakening of the apparent Ca^{2+}-binding strength of the recombinant protein is caused by a destabilization of its apo form.

There have been several other reports of the effect of additional residues at the N terminus on the native-state stability of recombinant proteins, and a comparison of these with the present results will provide insight into a rule relating to the effects of an extra methionine residue in the proteins. Hargrove *et al.* (1994) have observed that the recombinant apomyoglobin expressed in *E. coli* is less stable than the authentic protein. They have also shown that the N terminus of recombinant apomyoglobin contains an extra methionine residue and that the structure around the N terminus is rigid. Polyhistidine tags in the N and C-terminal regions of Arc repressor (Milla *et al.*, 1993, 1995) have little effect on the stability and folding of the protein, whereas the polyhistidine tags of CspA alter the folding behavior by interacting with the wild-type portion of the protein (Reid *et al.*, 1998). The X-ray crystallographic structures of the Arc repressor (Raumann *et al.*, 1994) and CspA (Goldstein *et al.*, 1990) have shown that the structure around the N-terminal residue in CspA is rigid, whereas that of Arc repressor is flexible. The N-terminal region of staphylococcal nuclease is flexible (Hynes & Fox, 1991), and it has been reported that a 19-residue pro-peptide in the N-terminal region of the nuclease does not significantly destabilize the N state of the recombinant protein (pro staphylococcal nuclease; Suciu & Inouye, 1996). Therefore, these studies together with the our study strongly suggest that when the structure around the N-terminal residue of a protein is rigid, the addition of extra residues at the N terminus destabilizes the N state of the protein. On the other hand, when the structure is flexible, the extra residues do not interfere with the native-state stability. From these experimental results, we can thus conclude that when the N-terminal region of residue, destabilizes the N state, but that when the N-terminal region is flexible, expression of the protein by *E. coli* does not interfere with the native-state stability.

Electric properties of authentic and recombinant α-lactalbumin

The results of the electrophoresis and ion-exchange chromatography show that the recombinant protein is more negatively charged than the authentic one. It is understood, however, that the side-chain of a methionine residue does not ionize at neutral pH, so that there is no difference in the number of ionizable groups between the authentic and recombinant proteins. In fact, our electrophoresis and ion-exchange chromatography data show that there is no difference in the electric charge between the proteins in the presence of 8 M urea. This means that some of the ionizable groups that have a pK_a near 7.0 experience a change in pK_a due to the structural folding of the protein. There are two such ionizable groups, the imidazole group of histidine and the N-terminal amino group, which have intrinsic pK_a values of 6.5 and 8.0, respectively. If we compare the X-ray structures of the two proteins, no significant differences are observed near the histidine side-chains. However, there is a noticeable difference in the structures around the N-terminal amino groups. The N-terminal amino group is hydrogen-bonded to the oxygen atom of Thr38 and is closer to the side-chains of Asp37 and Asp83 in the authentic protein (Figure 8), and both of these may increase the pK_a value of the N-terminal amino group through electrostatic interactions. A study of the pH-dependence of the unfolding transition of authentic bovine α-lactalbumin has shown that the N-terminal amino group of the protein has an abnormally high pK_a value ($pK_a = 8.9$) in the N state, which is normalized on unfolding from the N to the molten globule state (Kuwajima *et al.*, 1981).

It should also be mentioned that the ΔpK_a of the N-terminal group between the recombinant and authentic proteins leads to a difference in the native-state stability between the proteins, but this stability difference is expected to be much smaller than the $\Delta\Delta G_U$ estimated from equation (6) at pH 7.0. The stability difference ($\Delta\Delta G_U(\Delta pK_a)$) due to the ΔpK_a is known to be given by:

$$\Delta\Delta G_U(\Delta pK_a) = RT\ln[(1 - K_a(\mathrm{rec})/[H^+])/(1 - K_a(\mathrm{auth})/[H^+])] \qquad (6)$$

where K_a(rec) and K_a(auth) are the dissociation constants of the N-terminal amino groups of the recombinant and authentic proteins, respectively, and $[H^+]$ is the hydrogen-ion concentration (Tanford, 1970). If we assume that the pK_a(rec) and pK_a(auth) are 8.0 and 8.9, respectively, the above equation gives a $\Delta\Delta G_U(\Delta pK_a)$ of 0.06 kcal/mol at pH 7.0, and this is negligibly small compared with

N-terminal amino group reasonably interprets the differences in the electric properties between the proteins observed by electrophoresis and ion-exchange chromatography, but it is not sufficient for interpreting the stability difference between the proteins.

Folding of goat α-lactalbumin

Because the presence of the N-terminal methionine residue in the recombinant protein changes the thermodynamic stability of the native state, this system is useful for investigating the role of the N terminus in the folding of α-lactalbumin. We thus investigated the refolding and unfolding kinetics of the proteins by stopped-flow CD measurements. The results show that the rate of unfolding of the recombinant protein is faster than that of the authentic protein (Table 2), whereas the refolding rates are very similar in the two proteins (Table 3). This shows that the stability difference is caused by the enhanced unfolding rate of the recombinant protein, and this is interpreted in terms of the difference in the free energy of the unfolding transition ($\Delta\Delta G_U$) and the difference in the activation free energy ($\Delta\Delta G^{\ddagger}_U$) of unfolding. The $\Delta\Delta G^{\ddagger}_U$ is known to be given by the ratio of the unfolding rate constants as:

$$\Delta\Delta G^{\ddagger}_{U.} = -RT\ln\left[\frac{k_u(\mathrm{rec})}{k_u(\mathrm{auth})}\right] \qquad (7)$$

where k_u(rec) and k_u(auth) represent the unfolding rate constants for the recombinant and authentic proteins, respectively. Because k_u(rec) is 5.7 times larger than k_u(auth) at 5.4 M GdnHCl, $\Delta\Delta G^{\ddagger}_U$ is estimated to be 1.0 kcal/mol, and this value is nearly identical with the $\Delta\Delta G_U$ (0.8 kcal/mol) at the same concentration of the denaturant. Thus, the stability difference between the proteins can be fully interpreted in terms of the increase in the unfolding rate of the recombinant protein. This means that the structure around the mutation site, the N terminus in this case, has not yet been organized in the transition state of refolding in α-lactalbumin (Kuwajima *et al.*, 1989; Matouschek *et al.*, 1989; Serrano *et al.*, 1992). The folding initiation site of α-lactalbumin is thus not located in the N-terminal region of the protein. Previous studies have shown that the structure around the 6-120 disulfide bond and that around the B helix, both of which are involved in the α-domain of this protein, have not yet been organized in the transition state of refolding (Ikeguchi *et al.*, 1998; T. Y. *et al.*, unpublished data), while the structure around the Ca^{2+}-binding site is known to be already organized in the transition state (Kuwajima *et al.*, 1989). Our results thus provide further support for the proposition that the folding initiation site of α-lactalbumin is located at the interface between the α and β-domains, around the Ca^{2+}-binding site of the protein.

Materials and Methods

Chemicals

GdnHCl was of a specially prepared reagent grade for biochemical use from Nacalai Tesque, Inc. (Kyoto). The concentration of GdnHCl was determined from the refractive index at 589 nm with an Atago 3T refractometer (Pace, 1986). Cyanogen bromide (CNBr) was purchased from Nacalai Tesque Inc. (Kyoto). Authentic goat α-lactalbumin was prepared from fresh goat milk by the method described (Kuwajima *et al.*, 1980). A Resource^{TM}-Q anion exchange column was purchased from Pharmacia Biotechnology, Inc. (Sweden) and a μ BONDASPHERE 5 μ C4 300 Å reversed-phase column was supplied by Nihon Waters Ltd (Japan).

Expression and purification of recombinant goat α-lactalbumin

The expression system of goat α-lactalbumin and the procedures for the refolding and purification of the protein have been reported by Kumagai *et al.* (1990) and recently improved by Uchiyama *et al.* (1995) utilizing a T7 promoter (Studier & Moffatt, 1986). In brief, the protein expressed in *E. coli* BL21(DE3) as inclusion bodies was solubilized in 8 M urea containing 20 mM Tris-HCl (pH 8.0) and first purified using a DEAE-Sepharose FF column. The eluted protein was reduced by 50 mM dithiothreitol and dialyzed against 20 mM Tris-HCl (pH 8.0) at 4 °C to remove urea. Refolding of the reduced α-lactalbumin was performed as described (Sawano *et al.*, 1992), with slight modifications, in a solution containing 20% (v/v) glycerol, 20 mM Tris-HCl (pH 8.0), 1 mM $CaCl_2$, 6 mM glutathione, 0.6 mM oxidized glutathione, 3.3 μM α-lactalbumin at 15 °C for more than 20 hours. The refolding process was monitored by the appearance of a sharp peak on a reversed-phase HPLC chromatogram detected by UV-absorbance at 215 nm using a C4 column with a linear gradient elution of 28 %-52 % acetonitrile in the presence of 0.1 % (v/v) trifluoroacetic acid at a flow rate of 0.5 ml per minute. The refolded protein was then purified by DEAE-Sepharose FF and phenyl-Sepharose CL column chromatographies as described by Lindahl & Vogel (1984). Concentrations of authentic and recombinant goat α-lactalbumin were determined spectrophotometrically using an extinction coefficient of $E^{1\%}_{1\,\mathrm{cm}} = 20.1$ for both (Kuwajima *et al.*, 1980). No free cysteinyl residues were detected in the folded recombinant protein by thiol content analysis (Ellman, 1959; Riddle *et al.*, 1979).

Preparation of methionine-free recombinant goat α-lactalbumin

The methionine-free protein was prepared according to the method described by Kim *et al.* (1997) with slight modifications. Recombinant goat α-lactalbumin was dissolved in 70% (v/v) formic acid and treated with 100 mM CNBr (50-100-fold molar excess over the protein concentration) for 24 hours in the dark at room temperature. The cleaved product was diluted ten times with water and dialyzed against 10 mM HCl, then dialyzed against 10 mM Tris-HCl (pH 8.5) containing 1 mM $CaCl_2$. Finally, the protein solution was purified on a Q-Sepharose FF column, which had been equilibrated with 20 mM Tris-HCl (pH 8.5) containing 1 mM $CaCl_2$ and eluted with a linear gradient of NaCl from 0 M to 0.5 M. The mobilities and retention times of the eluted fractions were checked by native PAGE and anion-exchange HPLC, and compared with those of the authentic protein under the same conditions. The mass of the methionine-cleaved protein was determined by mass spectrometric analysis, and the removal of the N-terminal methionine residue was confirmed by the N-terminal sequence analysis. The concentration of the CNBr-cleaved protein was calculated using the same extinction coefficient as that given above.

Mass spectrometric analysis

Mass spectrometric analyses of the authentic, recombinant and methionine-free proteins were carried out by the MALDI-TOF-MS mass spectroscopic method. Sinapinic acid mix protein samples were used as the matrix, and the spectra were taken in Reflex (Bruker).

N-terminal sequence analysis

N-terminal sequencing of recombinant, authentic, and CNBr-cleaved proteins were carried out using an automated Applied Biosystem sequencer model 477a equipped with a model 120A on-line PTH amino acid analyzer. In this study we analyzed the first five residues in the proteins.

Equilibrium CD measurements

Equilibrium CD spectra were taken on a Jasco J-720 spectropolarimeter using an optical cuvette with a path length of 1.00 mm for measurements in the peptide region and 10.0 mm for measurements in the aromatic region. The CD spectra of the protein were measured in 50 mM sodium cacodylate, 50 mM NaCl (pH 7.0) containing 1 mM $CaCl_2$. The solutions for the GdnHCl-induced equilibrium unfolding studies were prepared in the same buffer containing various concentrations of GdnHCl. The mean residue ellipticity was calculated as a function of GdnHCl concentration at 25 °C by taking 113 as the mean residue mass. The protein concentration in the equilibrium measurements was 0.15-0.2 mg/ml.

The apparent fractional extent (F_{app}) of unfolding was calculated by:

$$F_{app} = \frac{\theta_{obs} - \theta_N}{\theta_U - \theta_N} \tag{8}$$

where θ_{obs} is the observed ellipticity, and θ_N and θ_U are the ellipticities in the native (N) and the fully unfolded (U) states, respectively. The θ_N and θ_U values are assumed to linearly depend on the GdnHCl concentration (C) as $\theta_N = \theta_1 + a_1 C$ and $\theta_U = \theta_2 + a_2 C$. The N state baseline was calculated from the ellipticity values between 0.5 and 2 M and between 0.4 and 1.8 M GdnHCl, and the U state baseline was from the values between 4.5 and 6.2 M and between 3.8 and 6.2 M GdnHCl for the authentic and recombinant proteins, respectively.

Kinetic measurements

Refolding and unfolding reactions of the authentic and recombinant proteins were induced by GdnHCl concentration jumps, which were performed by a stopped-flow CD apparatus (UNISOKU Inc., Japan) installed in the cell compartment of the J-720 spectropolarimeter sodium cacodylate at pH 7.0 and 25 °C. The dead time of the stopped-flow CD apparatus was 25 ms when a 4 mm cuvette was used. The concentration of the protein stock solution was about 1.5-2.0 mg/ml. The initial protein solutions before the concentration jump contained 1.0 M and 5.5 M GdnHCl for unfolding and refolding experiments, respectively. The diluent solution contained the same buffer (50 mM sodium cacodylate, 50 mM NaCl, and 1 mM $CaCl_2$, pH 7.0) and an appropriate concentration of GdnHCl. The two solutions were mixed with a mixing ratio of 1:10.

X-ray crystallographic studies

The crystal of recombinant goat α-lactalbumin was grown by the vapor diffusion method with a hanging drop in a chamber where the temperature was controlled at 20 °C. The data were collected by an automated area detector system, DIP2000, on an X-ray generator with a bent mirror system at 9.5 °C. Data processing and reduction was performed using DENZO and SCALEPACK programs (Otwinowski, 1993). The crystallographic data, the diffraction intensity statistics, and the refinement statistics are listed in Table 4. The crystal structure was solved on the basis of the model structure of baboon α-lactalbumin (Acharya *et al.*, 1989) by the molecular replacement method (Brünger, 1990) and was refined by a slow-cooling molecular-simulated annealing method in the X-PLOR 3.1 program suite (Brünger, 1992).

Theoretical estimation of $\Delta\Delta G_U$ between recombinant and authentic goat α-lactalbumin

The $\Delta\Delta G_U$ value was calculated by the method described by Oobatake & Ooi (1993). In this calculation, every atom was identified as belonging to one of seven atomic groups: aliphatic C, aromatic C, hydroxyl O, amide N, carbonyl C, carbonyl O, and sulphur S. In addition, the accessible surface area (ASA) of each atom in the N state (except hydrogen) was calculated by the method described by Richmond (1984) using the coordinates of the X-ray crystal structures. Because the N-terminal methionine residue is present only in the recombinant protein, the $\Delta\Delta G_U$ was assumed as a first approximation to be equal to the free energy change of unfolding (ΔG^u) of the methionine residue. For the ASA of atoms in the methionine residue in the U state, the values calculated by Shrake & Rupley (1973) were used. It was also assumed that the ΔG^u_h and ΔH^u_c are proportional to the change in the ASA ($\Delta\alpha_i$ for the *i*th atomic group) of the atoms on unfolding according to Oobatake & Ooi (1993). Thus:

$$\begin{aligned}\Delta G^u &= \Delta G^u_h + \Delta G^u_c\\ \Delta G^u_h &= \Sigma_i g_{i,h} \Delta\alpha_i\\ \Delta G^u_c &= \Delta H^u_c - T\Delta S^u_c\\ \Delta H^u_c &= \Sigma_i h_{i,c} \Delta\alpha_i\end{aligned} \tag{9}$$

where $g_{i,h}$ and $h_{i,c}$ are proportionality constants for the seven atomic groups. Although the change in the conformational entropy, ΔS^u_c, was also assumed to be proportional to the $\Delta\alpha_i$ values in the original Oobatake & Ooi (1993) method, this assumption may not be correct for the extra methionine residue of recombinant goat α-lactalbumin due to the rigid nature of this residue as the distance between the C^{ε} atom of the methyl side-chain of Met0 and the C^{δ} atom of Gln2 side-chain being 3.5 Å. Moreover, the N-terminal amino group in the recombinant protein is hydrogen-bonded with the carbonyl oxygen atom of the Gln39 side-chain. Thus the $-T\Delta S_c^u$ value (−5.9 kcal/mol) obtained from Table 8 of Oobatake & Ooi (1993) was employed (see Discussion).

The contribution of other residues to the $\Delta\Delta G_U$ value was also estimated by $\Sigma_i\ (g_{i,h}+g_{i,c})\Delta\alpha_i^N$, where $g_{i,c}$ is a proportionality constant and $\Delta\alpha_i$ is the difference in the ASA value of the ith atomic group between the authentic and recombinant proteins in the N state. Here, $-T\Delta S_c^u$ was assumed to be proportional to the change in the ASA values following the original Oobatake & Ooi (1993) method. The values obtained for Mol A and Mol B of the crystal structure of the authentic protein were averaged. Since Glu1 of the authentic protein is more exposed to solvent in the unfolded state, the difference in the ASA of the atoms of Glu1 between the authentic and recombinant proteins in the unfolded state ($\Delta\alpha_i^U$) was also taken into account for the estimation of $\Delta\Delta G_U$ as $\Sigma_i\ (g_{i,h}+g_{i,c})\ \Delta\alpha_i^U$ using the ASA values of Shrake & Rupley (1973). The contribution of the other residues to the $\Delta\Delta G_U$ thus estimated has been found to be less than 1 kcal/mol.

Protein Data Bank accession number

The coordinates have been deposited in the Brookhaven Protein Data Bank with accession number 1HMK.

Acknowledgments

The authors thank Mr K. Maki, Mr T. Makio, Mr T. Hayashi, and Mr M. Mizuguchi for their assistance in this work. This work was supported by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture and Science of Japan. T.K.C. is a postdoctoral Fellow of the Japan Society for the Promotion of Science.

References

Acharya, K. R., Stuart, D. I., Walker, N. P. C., Lewis, M. & Philips, D. C. (1989). Refined structure of baboon α-lactalbumin at 1.7 Å resolution. Comparison with C-type lysozyme. *J. Mol. Biol.* **208**, 99-127.

Acharya, K. R., Jingshan, R., Stuart, D. I., Philips, D. C. & Fenna, R. E. (1991). Crystal structure of human α-lactalbumin at 1.7 Å resolution. *J. Mol. Biol.* **221**, 571-581.

Adams, J. M. (1968). On the release of the formyl group from nascent protein. *J. Mol. Biol.* **33**, 571-589.

Arai, M. & Kuwajima, K. (1996). Rapid formation of a molten globule intermediate in refolding of α-lactalbumin. *Folding Des.* **1**, 275-287.

Aronsson, G., Martensson, L. G., Carlsson, U. & Jonsson, B. H. (1995). Folding and stability of the N-terminus of human carbonic anhydrase II. *Biochemistry*, **34**, 2153- 2162.

Balbach, J., Forge, V., Lau, W. S., van Nuland, N. A., Brew, K. & Dobson, C. M. (1996). Protein folding monitored at individual residues during a two-dimensional NMR experiment. *Science*, **274**, 1161-1163.

Brünger, A. T. (1990). Extension of molecular replacement: A new search strategy based on patterson correlation refinement. *Acta Crystallog. sect. A*, **46**, 46-57.

Brünger, A. T. (1992). *X-PLOR: Version 3.1. A system for crystallography and NMR*, Yale University Press, New Haven, CT.

Duverger, N., Murry-Brelier, A., Latta, M., Reboul, S., Castro, G., Mayaux, J. F., Fruchart, J. C., Taylor, J. M., Steinmetz, A. & Denefle, P. (1991). Functional characterization of human recombinant apolipoprotein AIV produced in *Escherichia coli*. *Eur. J. Biochem.* **201**, 373-383.

Ellman, G. L. (1959). Tissue sulfhydryl groups. *Arch. Biochem. Biophys.* **82**, 70-77.

Goldstein, J., Pollitt, N. S. & Inouye, M. (1990). Major cold shock protein of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **87**, 283-287.

Harata, K. & Muraki, M. (1992). X-ray structural evidence for a local helix-loop transition in α-lactalbumin. *J. Biol. Chem.* **267**, 1419-1421.

Hargrove, M. S., Krzywda, S., Wilkinson, A. J., Dou, Y., Ikeda-Saito, M. & Olson, J. S. (1994). Stability of myoglobin: a model for the folding of heme proteins. *Biochemistry*, **33**, 11767-11775.

Hiraoka, Y. & Sugai, S. (1985). Equilibrium and kinetic study of sodium- and potassium-induced conformational changes of apo-α-lactalbumin. *Int. J. Pept Protein Res.* **26**, 252-261.

Hynes, T. R. & Fox, R. O. (1991). The crystal structure of staphylococcal nuclease at 1.7 Å resolution. *Proteins: Struct. Funct. Genet.* **10**, 92-105.

Ikeguchi, M., Fujino, M., Kato, M., Kuwajima, K. & Sugai, S. (1998). Transition state in the folding of α-lactalbumin probed by the 6-120 disulfide bond. *Protein Sci.* **7**, 1564-1574.

Imoto, T., Yamada, H., Yasukochi, T., Yamada, E., Ito, Y., Ueda, T., Nagatani, H., Miki, T. & Horiuchi, T. (1987). Point mutation of alanine (31) to valine prohibits the folding of reduced lysozyme by sulfhydryl-disulfide interchange. *Protein Eng.* **4**, 333-338.

Ishikawa, N., Chiba, T., Chen, L. T., Shimizu, A., Ikeguchi, M. & Sugai, S. (1998). Remarkable destabilization of recombinant α-lactalbumin by an extraneous N-terminal methionyl residue. *Protein Eng.* **11**, 333-335.

Kataoka, M., Kuwajima, K., Tokunaga, F. & Goto, Y. (1997). Structural characterization of the molten globule of α-lactalbumin by solution X-ray scattering. *Protein Sci.* **6**, 422-430.

Katsumata, K., Okazaki, A., Tsurupa, G. P. & Kuwajima, K. (1996). Dominant forces in the recognition of a transient folding intermediate of α-lactalbumin by GroEL. *J. Mol. Biol.* **264**, 643-649.

Kim, S., Baum, J. & Anderson, S. (1997). Production of correctly folded recombinant [^{13}C, ^{15}N]-enriched guinea pig [Val90]- α-lactalbumin. *Protein Eng.* **10**, 455-462.

Kordel, J., Forsen, S. & Chazin, W. J. (1989). ^{1}H NMR sequential resonance assignments, secondary structure, and global fold in solution of the major (trans-Pro43) form of bovine calbindin D9k. *Biochemistry*, **28**, 7065-7074.

Kuhlman, B., Boice, J. A., Wu, W. J., Fairman, R. & Raleigh, D. P. (1997). Calcium binding peptides from α-lactalbumin: implications for protein folding and stability. *Biochemistry*, **36**, 4607-4615.

Kumagai, I., Takeda, S., Hibino, T. & Miura, K. (1990). Expression of goat α-lactalbumin in *Escherichia coli* and its refolding to biologically active protein. *Protein Eng.* **3**, 449-452.

Kuwajima, K. (1989). The molten globule state as a clue for understanding the folding and cooperativity of globular protein structure. *Proteins: Struct. Funct. Genet.* **6**, 87-103.

Kuwajima, K. (1996). The molten globule state of α-lactalbumin. *FASEB J.* **10**, 102-109.

Kuwajima, K., Nitta, S. & Sugai, S. (1980). Intramolecular perturbation of tryptophans induced by the protonation of ionizable groups in goat α-lactalbumin. *Biochim. Biophys. Acta*, **623**, 389-401.

Kuwajima, K., Ogawa, Y. & Sugai, S. (1981). Role of the interaction between ionizable groups in the folding of bovine α-lactalbumin. *J. Biochem. (Tokyo)*, **89**, 759-770.

Kuwajima, K., Mitani, M. & Sugai, S. (1989). Characterization of the critical state in protein folding. Effects of guanidine hydrochloride and specific Ca^{2+} binding on the folding kinetics of α-lactalbumin. *J. Mol. Biol.* **206**, 547-561.

Lindahl, L. & Vogel, H. J. (1984). Metal-ion-dependent hydrophobic-interaction chromatography of α-lactalbumins. *Anal. Biochem.* **140**, 394-402.

Luzzati, P. V. (1952). Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Crystallog.* **5**, 802-810.

Marcker, K. & Sanger, F. (1964). *N*-Formyl-methionyl-*S*-RNA. *J. Mol. Biol.* **8**, 354-360.

Matouschek, A., Kellis, J. T., Jr, Serrano, L. & Fersht, A. R. (1989). Mapping the transition state and pathway of protein folding by protein engineering. *Nature*, **340**, 122-126.

Miki, T., Yasukochi, Y., Nagatani, H., Fruno, M., Orita, T., Yamada, H., Imoto, T. & Horiuchi, T. (1987). Construction of a plasmid vector for the regulatable high level expression of eukaryotic genes in *Escherichia coli*: an application to over production of chicken lysozyme. *Protein Eng.* **4**, 327-332.

Milla, M. E., Brown, B. M. & Sauer, R. T. (1993). P22 Arc repressor: enhanced expression of unstable mutants by addition of polar C-terminal sequences. *Protein Sci.* **2**, 2198-2205.

Milla, M. E., Brown, B. M., Waldburger, C. D. & Sauer, R. T. (1995). P22 Arc repressor: transition state properties inferred from mutational effects on the rates of protein unfolding and refolding. *Biochemistry*, **343**, 13914-13919.

Miller, C. G., Strauch, K. L., Kurkal, A. M., Miller, J. L., Wingfield, P. T., Mazzei, G. J., Werlen, R. C., Graber, P. & Movva, N. R. (1987). N-terminal methionine-specific peptidase in *Salmonella typhimurium*. *Proc. Natl Acad. Sci. USA*, **84**, 2718-2722.

Mine, S., Ueda, T., Hashimoto, Y. & Imoto, T. (1997). Improvement of the refolding yield and solubility of hen egg-white lysozyme by altering the Met residue attached to its N-terminus to Ser. *Protein Eng.* **10**, 1333-1338.

Moerschell, R. P., Hosokawa, Y., Tsunasawa, S. & Sherman, F. (1990). The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine *in vivo*. Processing of altered iso-1-cytochromes *c* created by oligonucleotide transformation. *J. Biol. Chem.* **265**, 19638-19643.

Oobatake, M. & Ooi, T. (1993). Hydration and heat stab-

Otwinowski, Z. (1993). *DENZO: An Oscillation Data Processing for Macromolecular Crystallography,* Yale University, New Haven, C.T.

Pace, C. N. (1986). Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol.* **131**, 266-280.

Pfeil, W. (1998). Is the molten globule a third thermodynamic state of protein? The example of α-lactalbumin. *Proteins: Struct. Funct. Genet.* **30**, 43-48.

Pike, A. C., Brew, K. & Acharya, K. R. (1996). Crystal structures of guinea-pig, goat and bovine α-lactalbumin highlight the enhanced conformational flexibility of regions that are significant for its action in lactose synthase. *Structure*, **4**, 691-703.

Ptitsyn, O. B. (1995). Molten globule and protein folding. *Advan. Protein Chem.* **47**, 83-229.

Raumann, B. E., Rould, B. A., Pabo, C. O. & Sauer, R. T. (1994). DNA recognition by β-sheets in the Arc repressor-operator crystal structure. *Nature*, **367**, 754-757.

Reid, K. L., Rodriguez, H. M., Hillier, B. J. & Gregoret, L. M. (1998). Stability and folding properties of a model β-sheet protein, *Escherichia coli* CspA. *Protein Sci.* **7**, 470-479.

Richmond, T. L. (1984). Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. *J. Mol. Biol.* **178**, 63-89.

Riddle, P. W., Blakeley, R. L. & Zerner, B. (1979). Ellman's reagent: 5,5′-dithiobis (2-nitrobenzoic acid)-a reexamination. *Anal. Biochem.* **94**, 75-81.

Sawano, H., Koumoto, Y., Ohta, K., Sasaki, Y., Segawa, S. & Tachibana, H. (1992). Efficient in vitro folding of the three-disulfide derivatives of hen lysozyme in the presence of glycerol. *FEBS Letters*, **303**, 11-14.

Schulman, B. A. & Kim, P. S. (1996). Proline scanning mutagenesis of a molten globule reveals non-cooperative formation of a protein's overall topology. *Nature Struct. Biol.* **3**, 682-687.

Schulman, B. A., Kim, P. S., Dobson, C. M. & Redfield, C. (1997). A residue-specific NMR view of the non-cooperative unfolding of a molten globule. *Nature Struct. Biol.* **4**, 630-634.

Schultz, D. A. & Baldwin, R. L. (1992). *Cis* proline mutants of ribonuclease A. I. Thermal stability. *Protein Sci.* **1**, 910-916.

Serrano, L., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. *J. Mol. Biol.* **224**, 805-818.

Shrake, A. & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79**, 351-371.

Shimizu, A., Ikeguchi, M., Kobayashi, T. & Sugai, S. (1996). A synthetic peptide study on the molten globule of α-lactalbumin. *J. Biochem. (Tokyo)*, **119**, 947-952.

Shortle, D. (1996). The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J.* **10**, 27-34.

Studier, F. W. & Moffatt, B. A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189**, 113-130.

Suciu, D. & Inouye, M. (1996). The 19-residue pro-peptide of staphylococcal nuclease has a profound

Sugai, S. & Ikeguchi, M. (1994). Conformational comparison between α-lactalbumin and lysozyme. *Advan. Biophys.* **30**, 37-84.

Takeda, M. & Webster, R. (1968). Protein chain initiation and deformylation in *B. subtilis* homogenates. *Proc. Natl Acad. Sci. USA*, **60**, 1487-1494.

Tanford, C. (1970). Protein denaturation. C. Theoretical models for the mechanism of denaturation. *Advan. Protein Chem.* **245**, 4760-4769.

Uchiyama, H., Perez-Prat, E. M., Watanabe, K., Kumagai, I. & Kuwajima, K. (1995). Effects of amino acid substitutions in the hydrophobic core of α-lactalbumin on the stability of the molten globule state. *Protein Eng.* **8**, 1153-1161.

Vanderheeren, G. & Hanssens, I. (1994). Thermal unfolding of bovine α-lactalbumin. Comparison of circular dichroism with hydrophobicity measurements. *J. Biol. Chem.* **269**, 7090-7094.

Wilson, G., Hecht, L. & Barron, L. D. (1996). The native-like tertiary fold in molten globule α-lactalbumin appears to be controlled by a continuous phase transition. *J. Mol. Biol.* **261**, 341-347.

Wu, L. C. & Kim, P. S. (1997). Hydrophobic sequence minimization of the α-lactalbumin molten globule. *Proc. Natl Acad. Sci. USA*, **94**, 14314-14319.

Edited by **P. E. Wright**

(Received 22 June 1998; received in revised form 20 October 1998; accepted 22 October 1998)